# NO PAPER RIGHTS

---

# REIMPRESSION INTERDITE

Micromedia
ProQuest

20 Victoria Street
Toronto, Ontario M5C 2N8
Tel: (416) 362-5211
Toll free: 1-800-387-2689
Fax: (416) 362-6161

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de Technologie de l'information

ERB-1065

NRC·CNRC

# An Empirical Review of Software Process Assessments

Khaled El Emam and Dennis R. Goldenson
November 1999

Canada

NRC 43610

National Research
Council Canada

Conseil national
de recherches Canada

Institute for
Information Technology

Institut de Technologie
de l'information

# *An Empirical Review of Software Process Assessments*

Khaled El Emam and Dennis R. Goldenson
November 1999

Copyright 1999 by
National Research Council of Canada

Permission is granted to quote short excerpts and to reproduce figures and tables from this report, provided that the source of such material is fully acknowledged.

102-05395/

# An Empirical Review of Software Process Assessments[1]

**Khaled El Emam**
National Research Council, Canada
Institute for Information Technology
Building M-50, Montreal Road
Ottawa, Ontario
Canada K1A OR6
Khaled.El-Emam@iit.nrc.ca

**Dennis R. Goldenson**
Software Engineering Institute[2]
Carnegie Mellon University
Pittsburgh, PA 15213-3890
USA
dg@sei.cmu.edu

### *Abstract*

In the recent past there has been a rapid increase in the use of process assessments in the software industry. Such assessments play an important role in initiating and tracking process improvements. In addition, there has been a proliferation of best practice models that are the basis for the assessments. The purpose of this chapter is to review the empirical evidence that exists to date on the efficacy of process assessments, with a specific focus on the CMM for software and the emerging ISO/IEC 15504 international standard. The available evidence is quite extensive, and supports the claims that assessments can be an effective tool for process improvement. We also highlight gaps in the literature where further work needs to be done.

---

[1] ® CMM and Capability Maturity Model are registered in the U.S. Patent and Trademark Office.

℠ CMMI and IDEAL are service marks of Carnegie Mellon University.

[2] The Software Engineering Institute (SEI) is a federally funded research and development center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University.

# Table of Contents

**1 INTRODUCTION ........ 4**
1.1 The Stage Hypothesis ........ 4
1.2 The Second Wave ........ 5
1.3 Empirical Evaluation ........ 7
1.4 Scope and Caveats ........ 8
**2 SUMMARY OF REVIEW FINDINGS ........ 9**
**3 BACKGROUND ........ 11**
3.1 The Context of Assessments ........ 11
3.2 The Architecture of Best Practice Models ........ 13
**4 SOFTWARE PROCESS ASSESSMENT: MODELS AND METHODS ........ 17**
4.1 Coverage of Best Practice Models ........ 17
4.2 Assessment Methods ........ 17
4.3 Why do Organizations Perform Assessments ? ........ 19
4.4 Summary ........ 23
**5 SOFTWARE PROCESS IMPROVEMENT ........ 23**
5.1 The Order of Process Improvement ........ 23
5.2 SPI Experiences ........ 25
5.3 Creating Buy-in ........ 28
5.4 Factors Affecting Improvement Success ........ 29
5.5 The Cost of Software Process Improvement ........ 32
5.6 Summary ........ 35
**6 THE DIMENSIONS OF PROCESS CAPABILITY ........ 36**
**7 THE RELIABILITY OF PROCESS CAPABILITY MEASURES ........ 39**
7.1 Reliability Theory and Methods ........ 40
7.2 Internal Consistency ........ 43
7.3 Interrater Agreement ........ 46
7.4 Factors Affecting Reliability ........ 51
7.5 Summary ........ 54
**8 THE VALIDITY OF PROCESS CAPABILITY MEASURES ........ 55**

8.1 Types of Validity ........ 55
8.2 Predictive Validity Hypotheses ........ 58
8.3 Validation Approaches ........ 58
8.4 Main Effects ........ 60
8.5 Moderating Effects ........ 62
8.6 Diminishing Rates of Return ........ 62
8.7 Causality ........ 64
8.8 Summary ........ 65

**9 APPENDIX: AN OVERVIEW OF THE SPICE TRIALS ........ 65**

**10 APPENDIX: A REANALYSIS OF THE AFIT STUDY DATA ........ 67**

**11 ACKNOWLEDGEMENTS ........ 74**

**12 REFERENCES ........ 74**

# 1 Introduction

Over the last decade and a half there has been a rapid growth of interest by the software industry in Software Process Improvement (henceforth SPI). An industry to support SPI efforts has also been burgeoning. SPI can, and has, played an important role in achieving improvements in product quality and the ability to meet budget and schedule targets [118][131].

Two general paradigms to SPI have emerged, as described by Card [23]. The first is the analytic paradigm. This is characterized as relying on *"quantitative evidence to determine where improvements are needed and whether an improvement initiative has been successful".* The second, what Card calls the benchmarking[3] paradigm, *"depends on identifying an 'excellent' organization in a field and documenting its practices and tools".* The analytic paradigm is exemplified by the work at the Software Engineering Laboratory [165]. The benchmarking paradigm is exemplified by the CMM® for Software [157] and the emerging ISO/IEC 15504 international standard [64].[4] Benchmarking assumes that if a less-proficient organization adopts the practices of the excellent organization, it will also become excellent. Our focus in this chapter is on the benchmarking paradigm.

An essential ingredient of SPI following the benchmarking paradigm is a *best practice model*[5] (e.g., the Capability Maturity Model® - CMM - for software [157]). Such a model codifies what are believed to be best software engineering practices. By comparing their own processes to those stipulated in the best practice model, software organizations can identify where improvements in their processes should be made. The above process of comparison is known as a *Software Process Assessment* (henceforth SPA[6]).

A key feature of contemporary best practice models is that they group their software engineering practices into a series of 'stages' that an organization (or a process) should go through on its improvement path. By knowing which stage you are at, you can then identify the actions necessary to progress to the next higher stage. A SPA can be used to identify the organization's or process' current stage.

The objective of this chapter is to review the empirical evidence on the use of SPAs. The importance of such a review is perhaps best exemplified through a historical interlude. Hence, we first trace the path of Nolan's stage hypothesis. This is arguably the first 'stage' model, which initially appeared in print in 1973.

## 1.1 The Stage Hypothesis

One of the earliest stage models was Nolan's stage hypothesis [133]. Nolan observed the rise in the Information Systems (IS) budget of three firms, and interpreted this rise as following an S-shaped curve. From the points of inflection on these curves, he derived four stages that an IS organization goes through during its evolution: initiation, contagion, control, and integration. For example, the integration stage is characterized by the establishment of controls to allow the exploitation of computing without cost overruns. Planning is well established, Users are knowledgeable and capable in their uses of computing. Operations are rationalized, and economic analyses (e.g., cost/benefits analysis) are performed to justify and prioritize development work. And systems analysts are appropriately decentralized to user departments to encourage improved systems development. The stage hypothesis came to be regarded as a well-grounded empirical theory of IS organizational evolution [135], and became well entrenched in the Management Information Systems (MIS) discipline through its appearance in textbooks (e.g., [1][3][50][113]).

---

[3] This may also be called "experiential," since the term "benchmarking" is often used to describe the collection of quantitative data that can be used to describe the "state of the practice" among similar organizations in the software industry.

[4] Following the benchmarking paradigm involves using structured frameworks of best practice as identified by expert working groups and corroborated by external review. While such frameworks rely heavily on experience, they have in fact been based at least partially from the beginning on empirical analytic evidence as well.

[5] The model that is used as the basis for an assessment has been refered to differently over the years. For example, ISO/IEC 15504 refers to it as an "assessment model". The SW-CMM is refered to as a "reference model" by the SEI, which means something very different in the context of ISO/IEC 15504. To avoid confusion we shall use the term "best practice model" in this chapter.

[6] Here we use the term "SPA" in the general sense, not in the sense of an earlier SEI assessment method (which was also called a SPA). The successor to the SPA is the CBA IPI [48].

The term "assessment" is often reserved only for methods that are meant to be used by organizations for their own self improvement. The term "evaluation" is then used to refer to methods meant to be used by others for acquisition or source selection. However, "assessment" also is commonly used to cover both kinds of methods interchangeably. We use the term in its broader sense throughout this chapter.

Subsequent to the initial descriptive hypothesis, Nolan and his colleagues further refined it into a prescriptive model [81][82], with guidelines to aid the management of the IS function. Other work expanded the initial four-stage model into a six-stage model [134][135]: Initiation, Contagion, Control, Integration, Data Administration, and Maturity.

The importance and influence of Nolan's model also attracted a good amount of empirical evaluation. Researchers empirically evaluated the basic premises of the model as well as the accuracy of its predictions [7][126][86][45]. One review and interpretation of the empirical evidence [8] concluded that "most of the testable hypotheses have not been confirmed. [... ] the overall weight of the accumulated evidence to date in the evolution of IS research, is nonsupportive of the stage hypothesis. [... ] The general conclusion of all of the studies summarized here is that empirical support for the stage hypothesis is unconvincing. [... ] The empirical studies surveyed here indicate that the various maturity criteria do not reliably move together, or even always in the same direction, this refuting one of the requirements for claiming the existence of a stage theory." Some other authors have questioned the logical structure of the model, identifying a number of problems in its formulation [112].

Despite this evidence, the stage hypothesis was seen as a useful tool for the MIS discipline in that it promoted "a more organized approach to research on the subject", and it has value for its "conceptual contribution in the stage idea" [99]. Recently, some authors have argued that the observations embodied in Nolan's model were in fact essentially accurate, but that the model needed further refinement [88].

## 1.2 The Second Wave

Independent of Nolan's stage model, another stage model was developed in the mid-80's. Humphrey [100] initially described the software process maturity framework, which consists of five stages: Initial, Repeatable, Defined, Managed, and Optimizing. This is shown in Figure 1. The framework was based on earlier experiences at IBM [143]. It is both a descriptive and a guidance framework. It is descriptive of the "actual historical phases of evolutionary improvement of real software organizations" [100]. However, these software organizations can be considered best-in-class since the framework provides "a set of immediate priorities, once an organization's state in this framework is known". By addressing the identified priority items, you would then follow the path of the best-in-class organizations and gain the consequent performance improvements.

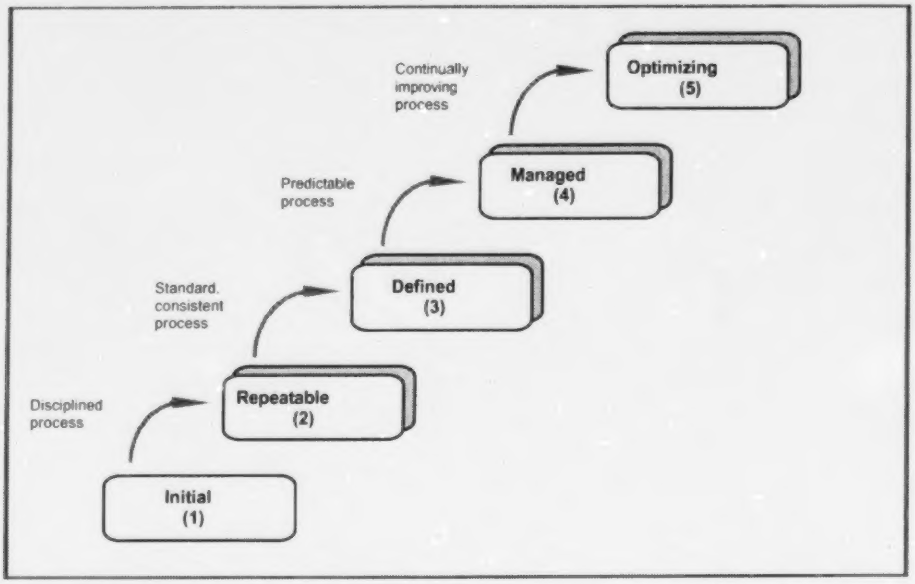


**Figure 1:** The stages of software process maturity according to the Humphrey framework (source [139]).

The software process maturity framework was not driven by academic research. Rather, it was meant to address a practical need: the increasing importance of software in DoD procurements and the need of all the [DoD] services to more effectively evaluate the ability of their software contractors to competently perform on software engineering contract" [101]. In fact, software procurement problems were not unique to the DoD. Drouin [44] describes similar problems encountered by the UK Ministry of Defence. Other large commercial procurers faced special difficulties, including the procurement of mission critical software before it is actually complete and that is expected to evolve to meet new requirements and continue to operate correctly for a number of decades [27]. In such cases, it is necessary to look at the capability of the supplier to deliver and maintain the product.

Humphrey and Sweet [101] operationalized the process maturity framework into a basic capabilty evaluation method that used a questionnaire to identify the "stage" at which an organization resided. The "stage" is known as the maturity level of the organization, taking on values from 1 (least mature) to 5 (most mature). The maturity level provides a simple quantification of the capability of a particular supplier, and can be used as input into the procurement process.

Early on, it was also realized that process assessments can be beneficial for SPI [100][30]. Process assessments were typically conducted to initiate and monitor improvement progress as part of larger SPI efforts. Hence, both capability evaluation of suppliers and process improvement became the two primary purposes for performing SPAs.

Humphrey's maturity framework evolved into the Capability Maturity Model for Software (SW-CMM) [7] [157]. The SW-CMM is the most well known best practice model today. But, by no means is it the only one. There has been an explosion of model production since the introduction of the SW-CMM. It is not possible to list them all, but some representative examples are the requirement engineering maturity model [158], a testing maturity model [20][21], a measurement maturity model [19], a systems engineering maturity model [156], the people capability maturity model [36], a maintenance maturity model [43], the Trillium model for telecommunications organizations [27], the SPR model [108], and Bootstrap [14][89][167]. Also, increasing in recognition and utility is the emerging ISO/IEC 15504 international standard [64][104], which defines requirements for models as well as an exemplar best practice model.

---

[7] A history of this evolution is provided in [138].

There have been some recent efforts at integrating the various models, most notably the Capability Maturity Model Integration (CMMI$^{SM}$) project[8]. The CMMI project is a joint effort of industry, government agencies, and the Software Engineering Institute at Carnegie Mellon University. Its product suite includes a framework for generating models, assessment methods, and related training materials. At this writing, a public review is currently underway of a model that integrates concepts from both software and systems engineering. A program of test and evaluation is about to commence in coordination with the international SPICE trials.

An important question that needs to be answered at the outset is how prevalent are assessments that use these models ? Herbsleb et al. [93] note that there are probably "thousands" of assessments that have been performed using the SW-CMM, and the resources expended on CMM-based improvement are in the "billions of dollars". A study by El Emam and Garro [69] estimated that between September 1996 and June 1998, approximately 1250 assessments using the emerging ISO/IEC 15504 standard were performed.[9] Based on the above numbers, it is reasonable to conclude that there are indeed many "thousands" of assessments being performed world-wide, and that these assessments represent a considerable investment of software engineering resources.[10]

## 1.3 Empirical Evaluation

So far we have established that SPAs are used for process improvement and procurement, and that they enjoy considerable usage worldwide by the software engineering community. But what empirical evidence exists to support and justify their use ? Given the experiences with Nolan's stage hypothesis, are we repeating the same pattern again ? Shouldn't we exercise prudence in empirically evaluating a technology first before deploying it so widely ?

Some of the comments that have appeared, and continue to appear, in the literature would make one think that the practical use of SPAs proceeds in an empirical vacuum. For instance, in the context of SW-CMM based assessments, Hersh[96] stated "despite our own firm belief in process improvement and our intuitive expectation that substantial returns will result from moving up the SEI [CMM] scale – we still can't prove it". Fenton [71] noted that evaluating the validity of the SEI's process maturity scheme is a key contemporary research issue. Jones, commenting on the SW-CMM effort at the SEI, stated that [110] "Solid empirical information is lacking on the quality and productivity levels associated with the SEI five levels of maturity". Very recently, Gray and Smith [87] render severe criticism of SPAs, mainly due to a presumptive lack of empirical evidence[11]. For example, they state "The whole CMM style approach is based on untested assertions and opinions, [... ] there is a lack of evidence to support the notions of repeatability and reproducibility of software process assessments. [... ] Currently software process assessment schemes are being applied and resultant process improvement plans being drawn up and acted upon without answers to the questions of repeatability and reproducibility being in place. [... ] the validity of the SEI's maturity model remains unproven. [..] If anyone states that existing process assessment schemes or the attempted world standard [ISO/IEC 15504] are based on a full understanding and sound theoretical underpinnings, let it go in one ear and out the other."

The above criticisms are undoubtedly rather severe. It is fair to say that at the outset when best practice models and SPAs were first introduced to the software engineering community, there was indeed little empirical evidence supporting their use. For example, in relation to SPAs, Pfleeger makes the point that [142] "many development organizations grab new, promising technologies well before there is clear evidence of proven benefit. For instance, the US Software Engineering Institute's Capability Maturity

---

[8] Further information about the CMMI project may be found at

[9] A point estimate of 1264 with a 90% confidence interval between 916 and 1895 was calculated using a capture-recapture model

[10] Although there are dissenting opinions. For example, [87] state that SPAs are not yet widely employed in the software industry (1998). Although, they do not present any evidence to support that argument.

[11] The authors make some claims about the general inefficacy of SPAs. However, none of these claims is supported by evidence either. This is surprising given that the main thrust of the article was the lack of empirical evidence supporting the use of SPAs

Model was embraced by many companies well before the SEI and others began empirical investigations of the nature and magnitude of its process improvement effects." Jones argues that [110] "the SEI Capability Maturity Model was deployed without any kind of serious evaluation or testing of the concept. [... ] The SEI assertions about quality and productivity were initially made without any kind of study or empirical evidence."

However, since their initial development and deployment, a considerable amount of empirical evaluative studies have in fact been performed. Especially when compared to other widely used software engineering technologies such as object-oriented techniques, the SPA field is remarkably rich with empirical evidence.

That empirical results lag the actual adoption of technology is not a problem unique to software engineering. In the allied discipline of MIS, this particular problem has been well recognized. Specifically, Benbasat and Zmud [9] state that rapid technological change "results in our chasing after practice rather than leading practice, and [this] typically leads to reporting results from (rigorous) studies involving new technologies years after the technology's acceptance (and, occasionally, rejection) by practice. [... ] Needless to say, pronouncements in the future about today's technological and associated business challenges are just not going to be considered relevant by most practitioners". Herbsleb [94] provides a further elaboration "If, for example, we are trying out a new technology that we believe has enormous potential, it may be a serious error for managers to wait until conclusive results are available. Management is the art of placing good bets, and as soon as a manger is convinced that the likelihood of significant benefits outweighs the risks and justifies the cost, managers should act."

Most of the empirical evidence did indeed come after wide initial adoption. When one considers the classical technology adoption stages [144], it becomes clear that empirical evidence serves the needs of early majority adopters who follow rather than lead, and who are willing to try new things only when they are demonstrated to be effective by others [142]. Innovators and early adopters need no such assurances. However, without the existence of innovators and early adopters, it is doubtful that realistic empirical evaluation studies can be performed at all.

Empirical evaluation serves two objectives. First, it can validate or refute claims made by developers about the costs and benefits of their technologies. In practice, usually one finds that some of the claims are confirmed under certain conditions, and are unsupportable under other conditions. Second, and perhaps more importantly, empirical evaluation can identify opportunities for enhancing the technology in the spirit of continuous improvement.

In this chapter we attempt to provide some answers to the concerns just voiced. We review the empirical evidence on SPAs to date, attempt to draw conclusions from existing studies, and distill what we have learned thus far. To be clear, those studies are not the last word on SPAs. More remains to be done to confirm and refine their conclusions, so we also attempt to identify promising areas for additional work.

## 1.4 Scope and Caveats

We limit our scope in this review largely to SPAs that are performed for the purpose of SPI rather than capability evaluation. While there are important differences between the two usages, much of the work on SPAs can be reasonably generalized to both SPI and capability evaluation, and a limited scope makes the review more focused.

In this review we focus mainly on the SW-CMM and ISO/IEC 15504, but we do provide some coverage of ISO 9000 due to its prevalence [169]. We are collectively most familiar with them, and a considerable amount of detailed information is publicly available about them. They have been used extensively worldwide. And a considerable number of empirical studies have evaluated them and their use.

At the outset we have to make clear two further caveats about this review. First, we only consider published works that we judge to be reasonably methodologically sound. Frequently it is heard at conferences or in meetings that company X has performed an empirical study that demonstrated Y. But the detailed results are not in the public record. Without access to a fully documented report, one cannot make an informed judgement about the weight of the evidence. A poorly done study lacks credence even if it supports a commonly held view. For the same reason, we deliberately have excluded studies that we judged to be based on questionable data or analytic methods. Second, there does exist a feedback loop whereby any weaknesses in best practice models and assessment methods that were identified in empirical studies were addressed in subsequent versions. The most recent versions of these models and methods in use today may actually be better than the available evidence suggests.[12]

# 2 Summary of Review Findings

Software Process Assessments (SPAs) are ubiquitous in today's software engineering industry. Encouraged by an increasing number of success stories, many organizations have begun to rely heavily on SPAs since their introduction in the mid 1980's. Some commentators continue to bemoan the adoption of a technology that has not been empirically evaluated. While this criticism may have been acceptable a few years ago, it is no longer defensible. An increasing number of empirical studies now exist. Of course, they do not demonstrate that SPAs are the solution to all of our problems. However the studies do highlight many strengths, as well as identify opportunities for continuing improvement.

In this chapter we provide a detailed review the empirical evidence about the efficacy of SPAs. What follows here is an overall summary of that review. The summary provides a comprehensive picture of the empirically based state of knowledge about SPAs, and also identifies promising areas for further exploration:

- At this writing, evidence exists that organizations perform assessments for the following reasons (see Section 4.3):
    - Sponsors expect a SPA to generate buy-in and create a climate for change within the organization.
    - Sponsors believe that process improvement based on the results of assessments will lead to bottom-line improvements in their projects and organizations.
    - Sponsors perceive SPAs as a valuable, objective measurement procedure.
    - Sponsors expect that a SPA will lead to the introduction of best practices within their organizations.
- The concept of “stages” found in best practice models is perceived to be useful by organizations performing SPAs. Further work needs to be performed, however, to clarify whether this is indeed the natural order in which historically successful software organizations improve (see Section 5.1).
- Approximately one third of respondents on SPI surveys report marked changes in their organizations as a consequence of their assessments. Organizations appear to have (unrealistically) high expectations from assessment-based SPI, and these need to be better managed. Despite the above, sponsors of assessments and assessment participants remain enthusiastic about the value of SPAs (see Section 5.2).
- Assessments increase “buy-in” for SPI for the participants in the assessments and the organizations' technical staffs. This is particularly encouraging since these are the individuals who are most likely to be initially skeptical about SPI, and to resist change from “yet another” management initiative (see Section 5.3).
- The most important factors that affect the success of SPI initiatives are (see Section 5.4):
    - Management commitment and support of SPI (e.g., management monitoring of the initiative, or making resources available).
    - Involvement of technical staff in the SPI effort.

---

[12] We assume that if empirical studies identify a strength in the various models and methods, the feature(s) in question will be retained. Therefore, we can safely claim that models and methods in use today are not likely to be worst than the evidence suggests.

  - Ensuring that staff understand the current software processes and their relationship to other business activities.
  - Clear SPI goals that are understood by the staff.
  - Tailoring the improvement initiatives.
  - Respected SPI staff (change agents and opinion leaders)
- In the chapter, we provide some general guidelines that may be useful in better managing expectations about the cost of SPI (see Section 5.5).
- The groupings of processes and practices within best practice models tend to be based largely on expert judgements made by model builders. Further empirical work needs to be performed in corroborating such judgments (see Section 6).
- A growing number of studies have empirically demonstrated the reliability of SPA results. Our ability to provide such evidence provides confidence in those results, and is particularly important given the high stakes of conducting process assessments (see Section 7). Studies of inter-rater agreement in ISO/IEC 15504 based assessments indicate that most ratings by independent teams are sufficiently reliable (at least 75%):
  - There exists evidence that more reliable assessments may prove to be less costly. When assessors agree in their initial rating judgments, the consolidation phase appears to require less consensus-building. Hence the consolidation progresses faster, resulting in an overall reduction in cost.
  - Early evidence suggests that the reliability of SPAs deteriorates with inexperienced assessors
  - Some systematic bias has been witnessed when assessment teams did not include both internal and external assessors. Mixed teams appear to be less prone to such unreliability.
  - The nature of the assessment method also appears to have a demonstrable impact on inter-rater agreement. One study found that rating judgments about lower capability levels were more reliable when they were deferred until later in the assessment when all of the evidence was collected and considered together.
- Further systematic empirical investigation is necessary, but a survey based study of assessment experts suggests a number of factors that may affect the reliability of assessments (see Section 7):
  - Clarity of the best practice model, and knowledge of it by the assessment team
  - The extent to which the assessment process is defined and documented
  - Amount of data that is collected during the assessment
  - Commitment to the assessment by both its sponsor and others in the assessed organization
  - Assessment team composition and stability
- By now there is ample evidence from a number of case studies that higher process capability is associated with improved product performance. More methodologically defensible studies of predictive validity also suggest that higher process capability tends to result in better performance. Similar results exist both at the project and organizational levels. (see Section 8)
- Further studies need to be performed to determine the rate of returns as an organization improves its processes according to current best practice models.

As we said at the outset, a considerable amount of empirical evidence already exists about the conduct and impact of software process assessments. However much more remains to be done, especially if we are to promote a more cumulative accumulation of knowledge:

- It is important that researchers use standardized instruments in their studies. For example, studies of predictive validity ought to reuse measures of process capability that were used in previous studies to allow comparability of the results. Without this, there will always be uncertainty as to what the body of knowledge really indicates.
- One of the vexing problems in correlational and observational studies of SPAs is the lack of appropriate sampling frames. This means that samples that are used for study most often are convenience samples. While this is driven by the constraints of a particular study, we encourage more effort to be placed on improved sampling. This will give us far greater confidence in the generalizability of our results.

# 3 Background

## 3.1 The Context of Assessments

The context of process assessment is depicted in Figure 2. This shows that process assessment provides a means of characterizing the current process capabilities of an organization. An analysis of the assessment results identifies the process strengths and weaknesses. For SPI, this would lead to an improvement initiative, which identifies changes to the processes in order to improve their capabilities. For capability determination, the assessment results identify whether the assessed processes meet some target capability. If the processes do not scale up to the target capability, then this may initiate an improvement effort.

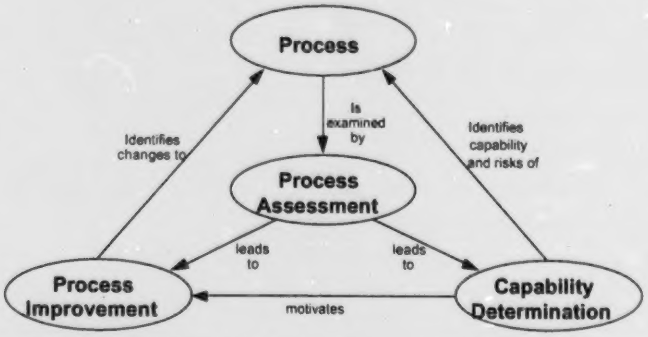


**Figure 2:** The context of process assessment (from [104]).

The performance of an assessment requires three different types of inputs. First is an assessment definition, which includes the identification of the assessment sponsor, the purpose and scope of the assessment, any relevant constraints, and identifying the assessment responsibilities (e.g., who will be on the assessment team, and who will be interviewed). It is also necessary to have an assessment method that describes the activities that need to be performed during an assessment.[13] Finally, an underlying best practice model is required. This model consists of the definitions of the processes that will be assessed, the assessment criteria, and a scheme to produce quantitative ratings at the end of the assessment.

Since our focus is SPI, process assessment's positioning within an overall SPI cycle can be seen in the IDEAL$^{SM}$ model [129] shown in Figure 3. It consists of five phases:[14]

I Initiating (the improvement program)

---

[13] The distinction between assessment definition and method is made in ISO/IEC 15504. Definition in that sense is part of "method" in CMM based appraisals.

[14] The description of IDEAL here is based on that given in [139]).

| | |
|---|---|
| D | Diagnosing (the current state of practice) |
| E | Establishing (the plans for the improvement program) |
| A | Acting (on the plans and recommended improvements) |
| L | Leveraging (the lessons learned and the business results of the improvement effort) |

The *Initiating* phase establishes the business reasons for undertaking a software process improvement effort. It identifies high-level concerns in the organization that can be the stimulus for addressing various aspects of quality improvement. Communication of these concerns and business perspectives is needed during the Initiating phase in order to gain visible executive buy-in and sponsorship at this very early part of the improvement effort.

The *Diagnosing* phase is used to build a common understanding of the current processes of the organization, especially the strengths and weaknesses of those current processes. It will also help identify priorities for improving your software processes. This diagnosis is based on a SPA.

The *Establishing* phase finalizes the strategy and supporting plans for the software process improvement program. It sets the direction and guidance for the next three to five years, including strategic and tactical plans for software process improvement.

The *Acting* phase takes action to effect changes in organizational systems that result in improvements in these systems. These improvements are made in an orderly manner and in ways that will cause them to be sustained over time. Techniques used to support and institutionalize change include defining software processes and measurements, pilot testing, and installing new processes and measurements throughout the organization.

The *Leveraging* phase completes the process improvement cycle. Lessons learned from the pilot projects and improvement efforts are documented and analyzed in order to improve the process improvement program for the future. The business needs that were determined at the beginning of the cycle are revisited to see if they have been met. Sponsorship for the program is revisited and renewed for the next cycle of software process improvement.

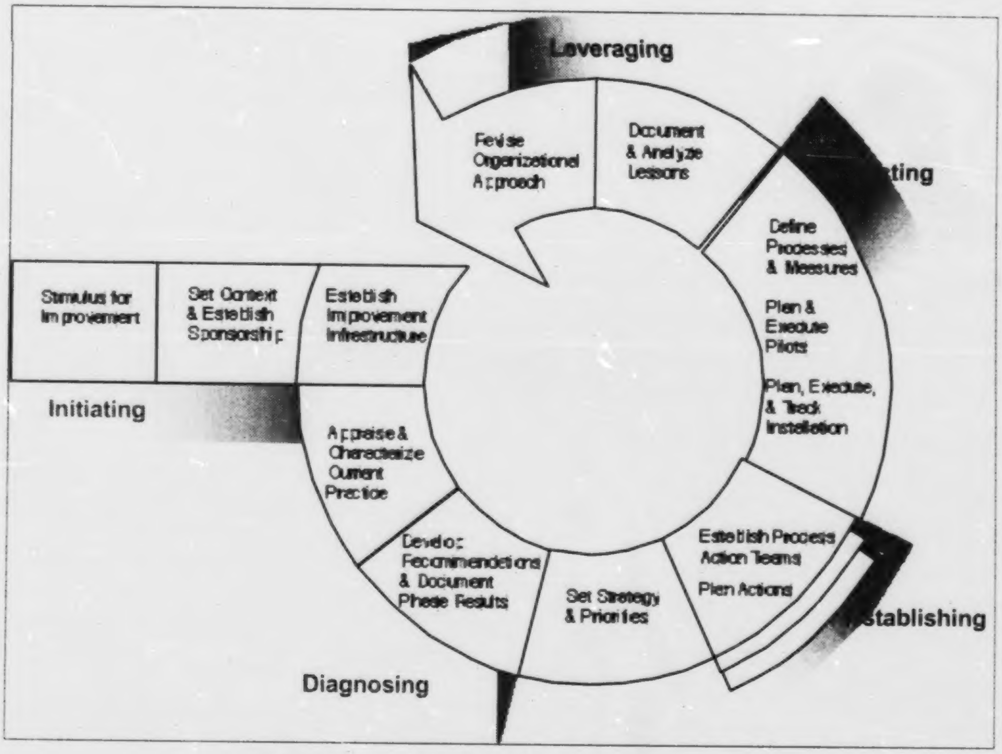

Leveraging
Revise Organizational Approach
Document & Analyze Lessons
ting
Define Processes & Measures
Plan & Execute Pilots
Plan, Execute, & Track Installation
Establish Process Action Teams
Plan Actions
tablishing
Set Strategy & Priorities
Develop Recommendations & Document Phase Results
Diagnosing
Appraise & Characterize Current Practice
Initiating
Stimulus for Improvement
Set Context & Establish Sponsorship
Establish Improvement Infrastructure

**Figure 3:** The SEI's IDEAL Model for SPI (source [139]).

## 3.2 The Architecture of Best Practice Models

The models on which process assessments are based can be classified under two different architectures[15] [137]. The architecture exemplified by the SW-CMM is known as a "staged" architecture. ISO/IEC 15504 specifies a "continuous" architecture.

The staged architecture of the SW-CMM is unidimensional since it defines a set of Key Process Areas (KPAs) at each maturity level (except level 1). The KPAs at each of the levels are shown in Figure 4. A maturity level is defined in terms of satisfaction of the goals of the KPAs within those levels. The ISO/IEC 15504 continuous architecture, on the other hand, is two dimensional as shown in Figure 5. One dimension consists of the processes that are actually assessed. As seen in the Figure, they are grouped into five categories on the Process dimension. The second dimension consists of the capability scale that is used to evaluate process capability (the Capability dimension). The same capability scale is used across all processes. (See Table 1 for further elaboration.)

The initial versions of ISO/IEC 15504 embodied the concept of a "process instance". A process instance is defined as a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner [64]. It is typical, but not necessary, that a process instance corresponds to a project. For example, if we were assessing the design process, then there would be one instance of the design process for each project within the organizational scope of the assessment. In this continuous architecture, a rating is made for each process instance. The ratings for each of the process instances are then aggregated to produce a rating for the design process of the organization. In the most recent version of ISO/IEC 15504 [104] it is no longer required to rate at the process instance, but rather one can produce a single rating for the process at the organizational level directly without explicit aggregation. The latter approach is more congruent with the rating scheme in the SW-CMM where the satisfaction of the KPA goals is rated at the organizational level [46].

---

[15] Current work in the CMMI project aims to harmonize the two architectures. CMMI models can have both staged and continuous representations.

The rating scheme used in SW-CMM based assessments also allows the aggregation of ratings across the KPAs to produce a single maturity level for the entire organization (within the scope of the assessment). For ISO/IEC 15504, the capability level rating is per process rather than for the whole of the organization.

Perhaps the biggest difference between the two architectures is that in a staged architecture the processes are ordered, while in a continuous architecture it is the capabilities that are ordered. For example, ISO/IEC 15504 does not require that a particular process must be defined at a level different than any other process, and any process can vary independently in its capability.[16] On the other hand the SW-CMM groups its KPAs into maturity levels, each of which represents increasing organizational capability.

| Level | Focus | Key Process Areas |
|---|---|---|
| **5<br>Optimizing** | ***Continual process improvement*** | **Defect Prevention<br>Technology Change Management<br>Process Change Management** |
| **4<br>Managed** | ***Product and process quality*** | **Quantitative Process Management<br>Software Quality Management** |
| **3<br>Defined** | ***Engineering processes and organizational support*** | **Organization Process Focus<br>Organization Process Definition<br>Training Program<br>Integrated Software Management<br>Software Product Engineering<br>Intergroup Coordination<br>Peer Reviews** |
| **2<br>Repeatable** | ***Project management processes*** | **Requirements Management<br>Software Project Planning<br>Software Project Tracking & Oversight<br>Software Subcontract Management<br>Software Quality Assurance<br>Software Configuration Management** |
| **1<br>Initial** | ***Competent people and heroics*** | |

**Figure 4:** Key Process Areas in the SW-CMM (source [139]).

---

[16] This distinction is not definitive, since the two dimensions in ISO/IEC 15504 are not completely orthogonal. There are strong links between some of the processes and capabilities.

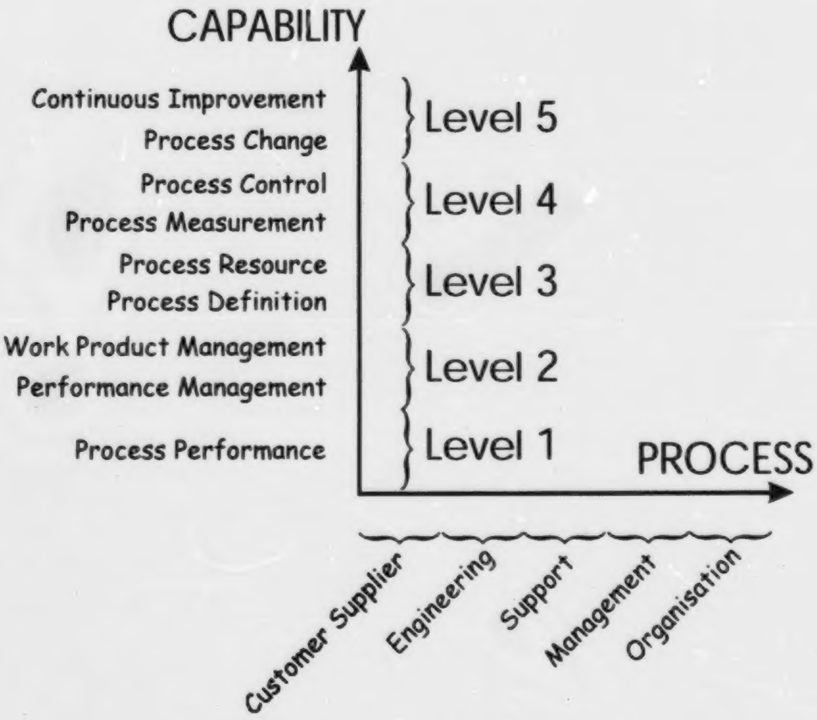


**Figure 5:** The architecture of ISO/IEC 15504.

In the remainder of this chapter we shall refer to the measurement of "process capability" to denote measurements using either of the architectures. This greatly simplifies the presentation since it reduces the amount of repetitive qualification that would otherwise be necessary. Moreover, one can in principle convert an ISO/IEC 15504 process capability profile into an overall organizational maturity rating through an appropriate aggregation. Hence we will refer to the finer grained measures here.

| ID | Title |
|---|---|
| Level 0 | **Incomplete Process**<br>There is general failure to attain the purpose of the process. There are no easily identifiable work products or outputs of the process. |
| Level 1 | **Performed Process**<br>The purpose of the process is generally achieved. The achievement may not be rigorously planned and tracked. Individuals within the organization recognize that an action should be performed, and there is general agreement that this action is performed as and when required. There are identifiable work products for the process, and these testify to the achievement of the purpose. |
| 1.1 | **Process performance attribute** |
| Level 2 | **Managed Process**<br>The process delivers work products of acceptable quality within defined timescales. Performance according to specified procedures is planned and tracked. Work products conform to specified standards and requirements. The primary distinction from the Performed Level is that the performance of the process is planned and managed and progressing towards a defined process. |
| 2.1 | **Performance management attribute** |
| 2.2 | **Work product management attribute** |
| Level 3 | **Established Process**<br>The process is performed and managed using a defined process based upon good software engineering principles. Individual implementations of the process use approved, tailored versions of standard, documented processes. The resources necessary to establish the process definition are also in place. The primary distinction from the Managed Level is that the process of the Established Level is planned and managed using a standard process. |
| 3.1 | **Process definition attribute** |
| 3.2 | **Process resource attribute** |
| Level 4 | **Predictable Process**<br>The defined process is performed consistently in practice within defined control limits, to achieve its goals. Detailed measures of performance are collected and analyzed. This leads to a quantitative understanding of process capability and an improved ability to predict performance. Performance is objectively managed. The quality of work products is quantitatively known. The primary distinction from the Established Level is that the defined process is quantitatively understood and controlled. |
| 4.1 | **Process measurement attribute** |
| 4.2 | **Process control attribute** |
| Level 5 | **Optimizing Process**<br>Performance of the process is optimized to meet current and future business needs, and the process achieves repeatability in meeting its defined business goals. Quantitative process effectiveness and efficiency goals (targets) for performance are established, based on the business goals of the organization. Continuous process monitoring against these goals is enabled by obtaining quantitative feedback and improvement is achieved by analysis of the results. Optimizing a process involves piloting innovative ideas and technologies and changing non-effective processes to meet defined goals or objectives. The primary distinction from the Predictable Level is that the defined process and the standard process undergo continuous refinement and improvement, based on a quantitative understanding of the impact of changes to these processes. |
| 5.1 | **Process change attribute** |
| 5.2 | **Continuous improvement attribute** |

**Table 1:** The levels of ISO/IEC 15504 defined, and their associated attributes (from [104]).[17]

---

[17] This is the capability scale that was in effect during the studies that are reported upon in this chapter. It has undergone some revisions since then though.

# 4 Software Process Assessment: Models and Methods

## 4.1 Coverage of Best Practice Models

Of course, best practice models do not necessarily cover all of the processes that are important for a particular organization. For example, in an early analysis of assessment data from 59 sites representing different business sectors (e.g., DoD contractor and commercial organizations) and different project sizes (from less than 9 peak staff to more than 100) [116], more than half of the sites reported findings that did not map into the KPA's of the software CMM.[18]

The lack of full coverage in any existing model most probably helps explain the current proliferation of models that include additional processes beyond the core of software engineering *per se*. One can argue that an organization should use the models that are most pertinent to its needs. However, that is far from simple in practice. First, not all models are equally well developed. For instance, by now much public guidance is available to assist in assessments and subsequent improvements based on the SW-CMM. Similar guidance is becoming available for the model specifications in ISO/IEC 15504 However the same is not true for all existing models. Furthermore, it is not evident that the reliability and validity results obtained with the SW-CMM and ISO/IEC 15504 are readily generalizable to all competing models (see later in this chapter).

It also is far from obvious that the results of assessments based on different models are directly comparable. For example, what if an organization performs an assessment using one model and obtains a rating of x, and performs another assessment using a different model that focuses on different processes, and obtains a rating of x-1? Which processes should the organization focus on first? Would it be those covered in the latter assessment because it resulted in a lower rating? One has no way of knowing *a priori*, since the measurement scales most probably are not comparable.

ISO/IEC 15504 may be helpful in this regard since it provides an overall framework that is intended to be applicable to many different models. The ISO/IEC 15504 documents define a set of requirements and the claim is that all ratings using best practice models that meet the requirements are comparable. However, this remains an empirical question. Similarly, the current CMMI effort aims to obtain consistent results for the same organizational and model scope of an assessment, regardless of the model representation that is used (staged or continuous).

## 4.2 Assessment Methods

An assessment method constitutes the activities that must be performed in order to conduct an assessment. There are many methods that are used in practice, most of them tailored to the needs of the assessed organizations. Examples of methods that have been described in the literature are [46][154]. Perhaps one of the best known methods is the SEI's CBA IPI, which is described in brief in [48], and with much greater detail in [46].

Some assessments that are performed with small organizations may last only a single day, whereas assessments in larger, high maturity organizations can take over two weeks of interviews, document reviews, and consolidation of findings.[19] A typical week long assessment would be organized as is shown in Figure 6.

---

[18] Note that revisions have been made to the CMM since that study, partly as a consequence of these findings.

[19] In an analysis of feedback data from CBA IPI assessments [47], about one third of lead assessors report spending three days on pre-onsite activities; another third report spending between four and eight days. Almost forty percent report devoting 5 days to on-site activities of the assessment; over an additional one half of the lead assessors said that they spend six or more days on on-site activities.

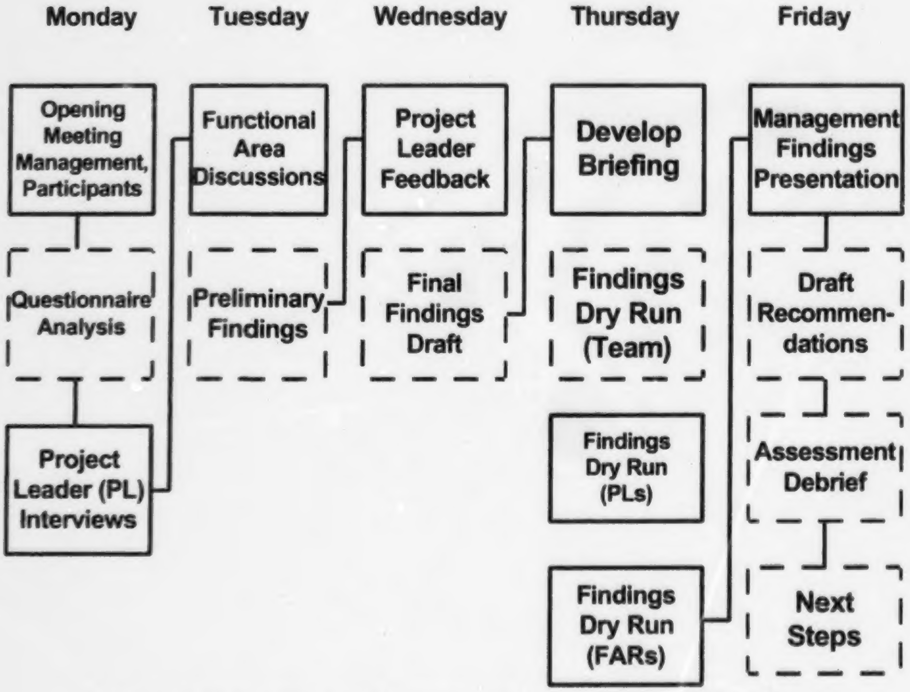


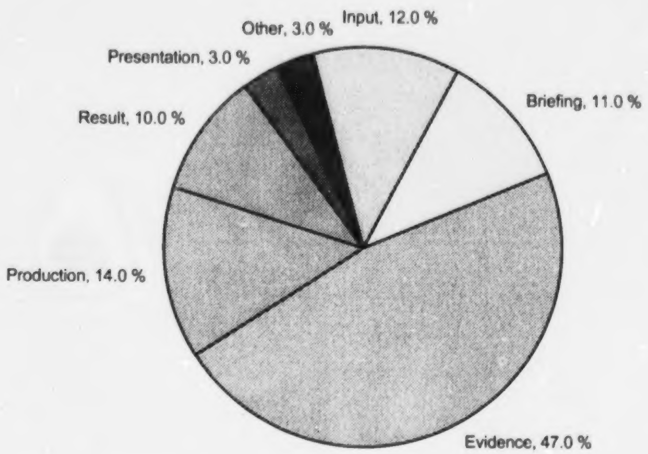
**Figure 6:** Organization of a typical week long assessment (from [49]).

The *CMM Appraisal Framework* (CAF) documents a set of requirements for assessment methods based on the SW-CMM [127]. The ISO/IEC 15504 documents contain another set of requirements (largely based on the CAF) [104]. Both sets of requirements are based on considerable expert knowledge and review. To our knowledge, however, there have been no systematic empirical investigations in this area

Aggregate data were collected as part of the SPICE Trials on the distribution of effort among the activities of a typical assessment. Assessors were asked to estimate the effort (in person-hours) spent on each of the following during the assessment:

- preparing the assessment input
- briefing the sponsor and/or organizational staff about the methodology to be used
- collecting evidence (e.g., reviewing documentation, interviewing organizational staff/management)
- producing and verifying ratings
- preparing the assessment results
- presenting the results to management.

The results are shown in Figure 7. Evidence collection is typically the most effort-consuming phase of an ISO/IEC 15504 assessment, followed by the production of the ratings. The initial phases of defining the assessment input and the initial briefing consume a non-negligible amount of effort as well.



**Cost Distribution**

**Figure 7:** Distribution of effort during assessments conducted in the SPICE Trials.

## 4.3 Why do Organizations Perform Assessments ?

It has been stated that SPAs for process improvement are performed for two reasons [48]:

- to support, enable, and encourage an organization's commitment to software process improvement, and
- to provide an accurate picture of the strengths and weaknesses of an organization's software processes

The former implies creating a climate for change within the organization[20]. In fact, it has been stated that [117] "The bottom line is that many [Process Improvement Program] activities succeed or fail based on the level of buy-in from the people involved". The latter implies obtaining a quantitative rating (or ratings) as a consequence of the assessment. Such ratings characterize the capability of the organization's processes.

Different authors will assign different weights to each of these two reasons for doing process assessments. For example, Dymond [49] asserts that the essence of a SPA is to "Agree on Process Issues". An assessment should generate consensus from management and the work force on the important process problems in the organization, and a commitment to address them. It is essentially a social ritual from that perspective. A prominent accident of assessments, according to Dymond, is the quantitative rating that is usually produced at the end.

Another posited reason for performing assessments is key customer or market pressure. This has been articulated in the case of the SW-CMM and DoD contractors [149] for the former. Also, some large procurers of software systems are including SPA's as an important ingredient in their supplier selection process [27][148], prompting suppliers to posses and demonstrate necessary process capabilities. For market pressure, a good example is ISO 9001, which often is required to do business in European Community nations [37][98].

---

[20] It would be expected that as the organization progresses successfully in its SPI effort, the effort required for creating a climate for change will be reduced.

Herbsleb and Grinter [95] describe a further benefit of SPAs based on a case study they performed in a large corporation. Namely, the assessment team was able to share information among the different development groups, across projects and business units. The assessment team acted as a central point for information about resolving problems raised from the results of the SPAs, and shared their knowledge across the corporation. Therefore, another reason for performing a SPA is to facilitate the acquisition and dissemination of information about best practices about which the assessment team is knowledgeable.

As can be seen above, a SPA may solve many problems simultaneously. The first question that one needs to ask is whether these problems are really important for the assessed organizations. For example, is it really important for the assessed organizations that a SPA create a climate for change ? How does this compare with the other reasons for performing an assessment ? Once we answer these questions in the affirmative, we can then evaluate how well SPAs meet these goals in practice.

An analysis performed by the SEI based on CBA IPI feedback data collected from assessment sponsors showed that over half of the sponsors stated that the primary goals of their assessments were either to monitor the progress of their existing software process improvement programs or to initiate new programs [47]. Furthermore, over a third of the sponsors said that validating an organization's maturity level was a primary goal of their assessment.

In an attempt to further provide some empirically grounded answers to the above questions, below we present a study of assessment sponsors that was performed during the SPICE Trials.[21]

In this study, assessment sponsors were asked about the reasons for performing the assessments in their organizations. The sponsors completed a questionnaire immediately after their assessments. Data are available from 70 assessments. All of the sponsors supported their respective organizations in performing at least one assessment using the emerging ISO/IEC 15504 standard. The questions are summarized in Table 2.[22]

---

[21] An overview of the SPICE Trials is provided in the appendix, Section 9.

[22] In our data set, some organizations were assessed more than once during the SPICE Trials. For some of these assessments, there were missing values on some of the 12 responses in Table 2. To deal with the missing values, we used the multiple imputation approach of Rubin [146]. Imputation techniques use auxillary models to estimate the missing values. In multiple imputation, these estimates are performed more than once, and then the analyses of the multiple data sets are combined.

There are two general approaches that one can take for imputation for this study:

- Assume that the reasons for conducting an assessment may vary for the same organization across different assessments. In this case, the unit of analysis is the assessment and the imputation approach allows for different response patterns for the same organization across different assessments.
- Assume that the reasons for conducting an assessment are constant for the same organization across multiple assessments. In this case the unit of analysis would be the organization, and the imputation approach would provide only one set of values for each organization.

Inspection of our data set indicates that the first assumption is more realistic. There were sponsors whom expressed different reasons for conducting an assessment across multiple assessments of the same organization. This may have been due to a different sponsor of the assessment. As another example, an organization may have conducted an initial assessment under customer pressure, but then the second assessment was driven by the organization itself after realising the benefits of assessments and wanting to track their SPI efforts.

| No. | Reason | Variable Name |
|---|---|---|
| 1 | Gain market advantage | ADVANTAGE |
| 2 | Customer demand to improve process capability | DEMAND |
| 3 | Improve efficiency | EFFICIENCY |
| 4 | Improve customer service | CUSTOMER |
| 5 | Improve reliability of products | PRODREL |
| 6 | Improve reliability of services in supporting products | SERVREL |
| 7 | Competitive/marketing pressure to demonstrate process capability | COMPETITIVE |
| 8 | Generate management support and buy-in for software process improvement | MANAGEMENT |
| 9 | Generate technical staff support and buy-in for software process improvement | TECHSTAFF |
| 10 | Establish best practices to guide organizational process improvement | BESTPRACT |
| 11 | Establish project baseline and/or track projects' process improvement | TRACKPROJ |
| 12 | Establish project baseline and/or track organization's process improvement | TRACKORG |

**Table 2:** The reasons why an assessment was performed. The wording of the question was "To what extent did the following represent important reasons for performing a software process assessment ?". The response categories were: "Don't Know", "Very Important", "Important", "Somewhat Important", "Not Very Important", "Not At All Important".

The "Don't Know" responses were treated as missing data. We therefore have a 5-point scale of importance (instead of the original six). There are two obvious alternatives for identifying the importance for each one of the above twelve reasons. The first is to dichotomize the five point scale into "Important" and "Not Important" categories. This was discounted, however, because we had difficulty deciding which category the response "Somewhat Important" should be in. This can be somewhat alleviated by dichotomizing around the median, for example. However, because the distributions that we obtained on these twelve questions tended to have sizeable clusters around the middle, the choice of the breakpoint as $\geq$ median or $\leq$ median would have had dramatic impact on the resulting dichotomous distribution. The second alternative, and the one that we opted for, was to use the mean.

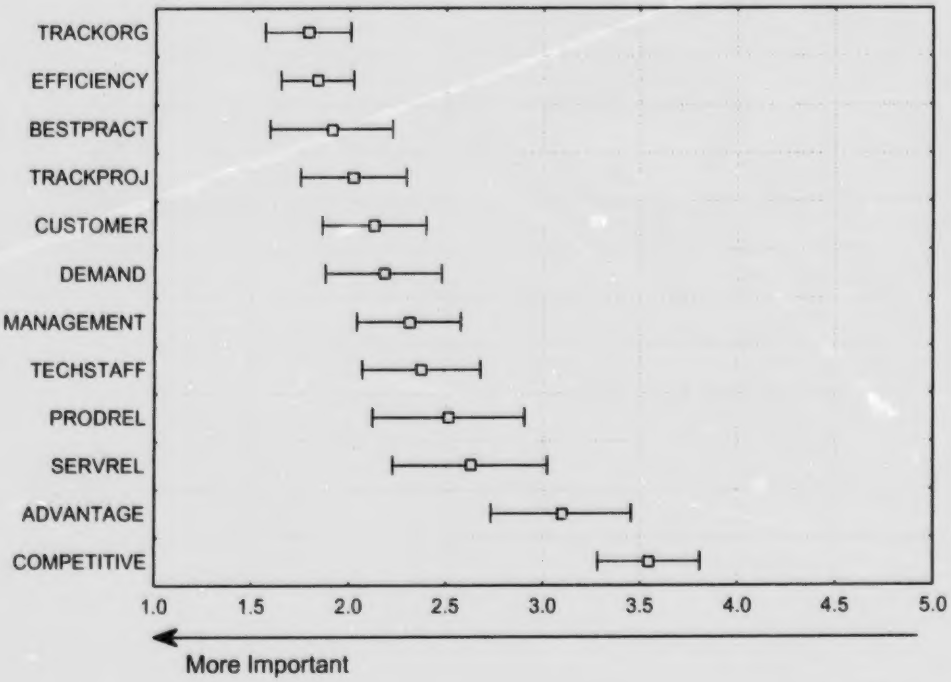


**Figure 1:** Mean of importance scores with 95% confidence interval. Low scores indicate higher importance, and the higher scores indicate less importance.

The range plot in Figure 1 shows the mean for each one of the reasons and the 95% confidence interval. This means that if we were to repeat this study a large number of times, and each time calculate the confidence interval, this interval will contain the population mean 95% of the time. To test a hypothesis of "Indifference" (i.e., that the respondent did not think that the reason was neither important nor not important for conducting an assessment), then we can use the tails of the confidence interval. Indifference can be defined as a response of "Somewhat Important", which has a value of 3. If the tails cross the value of three, then there is evidence, at a two-tailed $\alpha$ level of 0.1, that the mean response for that question is not different from 3. Only the SERVREL and ADVANTAGE variables demonstrate indifference.

Note first the reasons that were perceived *not* to be important. Clearly competitive/marketing pressure to demonstrate process capability (ADVANTAGE) was not a reason for performing assessments. This may be a reflection of the fact that those organizations that participated in the SPICE Trials are innovators and early adopters, and therefore their priorities may be different from the general population of organizations.

The sponsors also exhibited indifference on gaining market advantage and improving the reliability of services in supporting products as reasons for performing an assessment.

The most important reasons these sponsors chose for conducting their process assessments were to establish process capability baselines and/or track progress with project and organizational process improvement, to improve efficiency, and customer service, and to identify best practices for process improvement. The importance that the sponsors assigned to establishing capability baselines clearly indicates that they tended to recognize that assessments are in fact an important measurement procedure. Improving efficiency and customer service indicates that these sponsors believed that SPI based on the assessment would in fact provide tangible benefits to their projects. The identification of best practices is consistent with the conclusions of Herbsleb and Grinter [95], but also likely indicates that the sponsors expected the good practices embodied in the models would be transferred to their organizations.

In the middle range were the need to generate support and buy-in for process improvement among the technical staff and management, and customer demands for improving process capability. Again, these are consistent with the two basic reasons for performing assessments, namely to build support for process improvement as well as to accurately measure organizational capability.

An interesting issue to consider is whether there are differences in the reasons for performing assessments between small and large organizations. We dichotomize this IT staff size into SMALL and LARGE organizations, whereby small is equal to or less than 50 IT staff. This is the same definition of small organizations that has been used in a European project that is providing process improvement guidance for small organizations [163].

We computed the differences between small and large organizations in the reasons why they performed assessments. We found no discernable difference among the two organizational sizes. Therefore we can conclude that the reasons for performing SPAs are consistent irrespective of organizational size.

## 4.4 Summary

In this section we have provided an overview of process assessments for SPI, and have summarized the models and methods that are used to conduct them. Based on the results of a recent survey, we also described, , the reasons sponsors support process assessments in their organizations. These results provide a concrete set of criteria for evaluating SPAs to see if they meet their expectations. Specifically:

- Sponsors expect a SPA to generate buy-in and create a climate for change within the organization.
- Sponsors believe that process improvement based on the results of assessments will lead to bottom-line improvements in their projects and organizations.
- Sponsors perceive SPAs as a measurement procedure.
- Sponsors expect that a SPA will lead to the introduction of best practices within their organizations.

In the remainder of this chapter we review existing empirical evidence about how well SPAs in fact meet those criteria.

# 5 Software Process Improvement

In this section we review the studies performed on assessment-based SPI, how and why it succeeds.

## 5.1 The Order of Process Improvement

One of the basic tenets of all models on which process assessments are based is that the implementation of process improvements should follow a specific path, and this is the path stipulated in the best practice model. The existence of a path serves an important purpose, as articulated by Paulk [138] "The advantage of the maturity levels is that they provide clear priorities, which provide guidance for selecting those few improvement activities that will be most helpful if implemented immediately. This is important because most software organizations can only focus on a few process improvement activities at a time". Indeed, surveys of sponsors and assessors consistently find that they believe that feature to be a useful one, since it provides concrete guidance about the next process improvement steps that one should follow. In addition, the improvement steps are small and manageable. For example, 93% of the sponsors of ISO/IEC 15504 assessments who took part in the SPICE Trials agreed or strongly agreed with the statement that "The assessment provided valuable direction about the priorities for process improvement in the organization" [56]. In a survey of individuals whose organizations were assessed using the SW-CMM it was reported that over 80% believe that the "CMM provides valuable direction about the order in which process improvement should be made" [83].

However, an obvious question that one may ask is what is the basis for such an ordering of practices ? The logic of this sequencing is that this is the natural evolutionary order in which, historically, software organizations improve [100], and that practices early in the sequence are prerequisite foundations to ensure the stability and optimality of practices implemented later in the sequence [157].

The little evidence that does exist is not pertinent for contemporary models, but nevertheless fails to support the argument that earlier models embody the natural evolution of software engineering practices. In particular, one study investigated whether the maturity path suggested by Humphrey and Sweet's process maturity framework [101] follows a natural evolutionary progression [42].[23] Their analysis was based on the basic idea that questions representing maturity levels already passed by organizations would be endorsed (i.e., answered yes) while items representing maturity levels not reached would fail. Their results did not support the original maturity path and led the authors to suggest that the original model seemed "arbitrary" in its ordering of practices and is "unsupported". The first five levels of the alternative maturity model that they empirically derived are shown in Figure 2. Of course, further studies are necessary to confirm this alternative model, but at least it enjoys some empirical support thus far.

It should be noted that the ordering shown in Figure 2 does not come solely from organizations or projects that were successful. So it is not known whether this empirically derived ordering actually represents a path to success or to failure. Therefore, in reality we, as a community, still do not have systematic evidence as to the natural ordering of practices as followed by successful organizations and projects.

There are three implications of this. First, as of now, we still cannot make strong claims that best-practice models capture the true evolution of the implementation of software engineering practices in organizations. Second, that the current ordering of practices (or capability) in best practice models may still represent a logical ordering. This means that it is logical that some practices occur after others. It is also plausible that another ordering would be just as good and lead you to the same end state. Third, it is prudent to take the prescribed path in an best practice model as a suggestion rather than as gospel. If an alternative path or a modified path also makes business sense for a particular organization or project, then there is no compelling reason not to adopt it. In fact, in a report of SPI based on the CMM [18] it is noted that *"Business demands often necessitate improvements in an order which is counter to the CMM."* In that particular case, the organization initiated some process improvements that were not necessarily congruent with their next level of maturity, but driven by their business objectives.

---

[23] [23] Another unpublished study examined the ordering of capabilities in ISO/IEC 15504. However, this was done with organizations that were participating in the SPICE Trials, and therefore they were *a priori* predisposed to follow the stipulated path. Furthermore, the assessors may be predisposed in a manner that is consistent with the ordering of attributes in the capability dimension. As expected, the results provided a strong indication that organizations follow the 15504-stipulated order of achieving capability. Ideally, such studies should be done with organizations who are not already familiar with, or already using, a particular model to guide their improvement efforts.

**Level 1: Reviews and Change Control**

- Is a mechanism used for controlling changes to the code? (Who can make changes and under which circumstances?) (L2)
- Are internal software design reviews conducted? (L3)
- Are software code reviews conducted? (L3)
- Is a mechanism used for controlling changes to the software requirements? (L2)

**Level 2: Standard Process and Project Management**

- Is a mechanism used for controlling changes to the software design? (L3)
- Does the software organization use a standardized and documented software development process on each project? (L3)
- Do software development first line managers sign off on their schedules and cost estimates? (L2)
- Is a formal procedure used in the management review of each software development prior to making contractual commitments? (L2)
- Is a formal procedure used to produce software development schedules? (L2)
- Are formal procedures applied to estimating software development cost? (L2)
- Is a mechanism used for managing and supporting the introduction of new technologies? (L4)

**Level 3: Review Management and Configuration Control**

- Are the action items resulting from code reviews tracked to closure? (L3)
- Are the actions items resulting from design reviews tracked to closure?(L3)
- Are the review data gathered during design reviews analyzed? (L4)
- Is there a software configuration control function for each project that involves software development? (L2)
- Are code review standards applied? (L4)
- Is a formal procedure used to make estimates of software size? (L2)
- Is a mechanism used for periodically assessing the software engineering process and implementing indicated improvements? (L4)

**Level 4: Software Process Improvement**

- Are analyses of errors conducted to determine their process related causes? (L4)
- Is a mechanism used for ensuring compliance to software engineering standards? (L3)

**Level 5: Management of Review and Test Coverage**

- Are design and code review coverages measured and recorded? (L4)
- Is test coverage measured and recorded for each phase of functional testing? (L4)

**Figure 2:** Empirically derived maturity model (first 5 levels only). The levels in parentheses refer to the levels in the original Humphrey and Sweet framework [101]. This table is based on the results in [42].

## 5.2 SPI Experiences

There does exist evidence that SPI efforts succeed and bring benefits. However, not all organizations that attempt SPI based on an assessment are successful. Below we review the results of two studies on the success of assessment-based SPI initiatives.

### 5.2.1 The CMM Study

In an SEI survey [83], a sample of representatives from 61 assessments (in different organizations) were contacted. These assessments were conduced at least one year prior to the survey. Three respondents per organization were sought: (a) the project level software manager most knowledgeable about the assessment, (b) the most knowledgeable and well-respected senior developer or similar technical person available, and (c) an organizational level SEPG manager, or someone with equivalent responsibilities, if such a person existed. In total, responses from 138 individuals representing 56 of the 61 organizations were returned. This gives an average of about 2.5 responses per organization. The pooled observations from all three roles were used for the analysis, giving a total of 138 observations.

Over two-thirds of the respondents reported that their organization’s SPI efforts were largely determined by the findings and recommendations that were raised in the assessment. However, when asked “How successfully the findings and recommendations [from the assessment] have been addressed ?”, 70% of the respondents said “moderate”, “limited”, or “little if any”. Over one-fourth said that the recommendations resulting from their assessments proved to be too ambitious to accomplish in a reasonable period of time, 26% acknowledged that “nothing much has changed” since the assessment, 49% said that there “has been a lot of disillusionment over the lack of improvement”, 42% said that process improvement has been overcome by events and crises and that other things have taken priority, 72% report that process “improvement has often suffered due to time and resource limitations”, 77% say that process improvement has taken longer than expected, and 68% reported that it has cost more than they expected.

The above summary of the results indicate a number of things:

- SPI is difficult, with many organizations not being able to address the assessment recommendations and to demonstrate tangible changes to their processes as a consequence of the assessments.
- One potential reason can be seen from the responses, that there is not sufficient commitment of resources and energy within the organizations to make SPI happen.
- Also, it is clear that the expectations of organizations need to be managed better, in terms of the cost and time to make SPI happen.

However, 84% of the disagreed or strongly disagreed with assertions that software processes have become more rigid and bureaucratic or that it has become harder to find creative solutions to difficult technical solutions, and only 4% said that the assessments had been counter-productive and that the progress of SPI had actually worsened since their assessments.

### 5.2.2 The SPICE Trials Study

We summarize the results of a follow-up study of organizations that participated in an ISO/IEC 15504 assessment [65]. In this particular study, sponsors of assessments in the SPICE Trials were administered a questionnaire approximately one year after the assessment to determine their perceptions on the extent to which the assessment influenced their SPI efforts. This study was performed in the context of the SPICE Trials (an overview of the SPICE Trials is provided in the appendix, Section 9)

For the description of sponsors’ perceptions, we used percentages of respondents who are supportive (as opposed to critical) of their experiences with assessment-based SPI. For example, assume that a question asked the respondents to express their extent of agreement to the statement “The assessment was well worth the money and effort we spent; it had a major positive effect on the organization”, and that it had the following four response categories: “Strongly Agree”, “Agree”, “Disagree”, and “Strongly Disagree”. As shown in Figure 3, the “Strongly Agree” and “Agree” responses would be considered supportive of assessment-based SPI, and the “Disagree” and “Strongly Disagree” responses would be considered to be critical of assessment-based SPI.

| **Supportive Responses** | **Critical Responses** |
|---|---|
| *Strongly Agree*<br>*Agree* | *Disagree*<br>*Strongly Disagree* |

**Figure 3:** Scheme for defining supportive and critical responses.

We received a total of 18 responses to our questionnaire. However, some of the assessments were done too recently for any accurate information about the progress of SPI to be collected. Therefore, we excluded all observations that were conducted less than 30 weeks before the response time. This left us with data from 14 valid assessments and subsequent SPI efforts. The variation of elapsed time since the assessment is given in Figure 4. This indicates that the organizations from which we have data have conducted their assessments from 44 to 90 weeks before responding to the questionnaire. This provides sufficient time for SPI efforts to have started and for some progress to have been made.

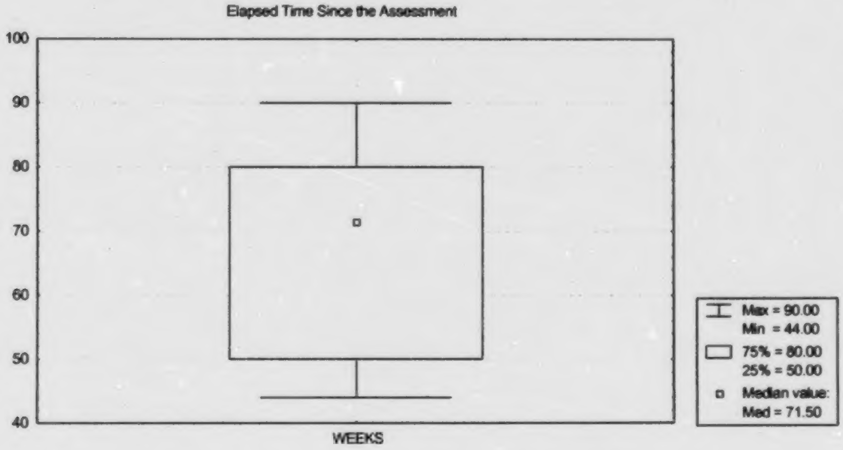


**Figure 4:** Distribution of elapsed time since the assessment.

| No. | Question | Percentage Supportive |
|---|---|---|
| (1) | Software processes have become more rigid and bureaucratic; it is harder to find creative solutions to technical problems | (14/14) = 100% |
| (2) | Because of the assessment, we have neglected other important issues facing the organization | (14/14) = 100% |
| (3) | The assessment was counter-productive; things have gotten worst | (13/14) = 93% |
| (4) | There has been a lot of disillusionment over the lack of improvement | (10/13) = 77% |
| (5) | Process improvement is costing more than we expected | (6/13) = 46% |
| (6) | The assessment was well worth the money and effort we spent; it had a major positive effect on the organization | (5/14) = 36% |
| (7) | Nothing much has changed since the assessment | (4/14) = 28% |
| (8) | Process improvement was overcome by events and crises; other things took priority | (3/14) = 21% |
| (9) | Process improvement is taking longer than expected | (2/13) = 15% |
| (10) | Process improvement has often suffered due to time and resource limitations | (1/14) = 7% |
| (11) | Process change has been easier than we expected | (0/14) = 0% |

**Figure 5:** Percentages of supportive and critical responses on the assessment based SPI.

Our results on the perceptions about assessment based SPI are summarized in Figure 5. Counter to what has sometimes been heard, all of the sponsors do not believe that software processes have become more bureaucratic due to their SPI efforts (see question 1). Neither do they believe that other important issues facing the organization have been neglected (see question 2). Almost all of the sponsors (93%)

do not agree with the statement that things have deteriorated (see question 3). Almost three quarters of the respondents (77%) disagree with the statement that there has been disillusionment due to a lack of improvement (see question 4).

However, a sizable group of respondents (54%) believe that SPI is costing more than they have anticipated (see question 5). Approximately three fifths do not believe that the assessment has had a major impact on the organization (see question 6). This may be due to there not being sufficient time since the assessment for SPI to have taken root, or due to the organizations not being able to act on the recommendations and findings from the assessment. To deal with this issue we can consider the recommendation that it can be very important to make some "quick wins" to ensure that the assessment is seen to have had an impact and to maintain momentum.

This is further reinforced by the finding that only 28% disagree with the statement that nothing much has changed since the assessment (see question 7). It is interesting to note that 79% believe that SPI was overcome by events and crises (see question 8), indicating potentially that the organizations were not in a state that is ready for long term SPI initiatives (i.e., there were other more important things to attend to). Some further reasons are forthcoming from the following responses: 85% believe that SPI is taking longer than expected (see question 9), and not surprisingly 93% believe that SPI suffered due to time and resource limitations (see question 10). None of the respondents believe that SPI has been easier than expected (see question 11).

Two general problems emerge from the above descriptive statistics. First, that expectations of organizational sponsors may not have been managed optimally, given that many believe that SPI costs more than expected, and is taking longer than expected. Second, in many cases insufficient resources were made available for SPI, and insufficient priority was given to SPI. This can lead us to the conclusion that an SPI effort must be treated as a project in its own right with a plan, resources and commitment. However, these problems do not seem to have dampened the enthusiasm within the organizations for assessments and assessment based SPI.

### 5.2.3 Summary

The results of the two studies above demonstrate a remarkable consistency:

- Approximately two thirds of respondents on SPI surveys do not report marked changes in their organizations as a consequence of their assessments: 70% of the SW-CMM survey respondents state that they had "moderate", "limited" or "little if any" when asked about how successful they were in addressing the recommendations from the assessment, 64% of the SPICE Trials survey respondents "disagree" or "strongly" disagree with the statement that the assessment had a major positive effect on the organization.[24]
- Organizations seem to have (unrealistically) high expectations from assessment-based SPI, and these need to be managed better.
- Despite the above, sponsors of assessments and assessment participants still remain enthusiastic about SPAs.

It is therefore important to consider the actual costs of SPI to provide more realistic expectations (see Section 5.5), and identify the critical success factors to make SPI happen to ensure that more organizations succeed in their assessment based SPI efforts (see Section 5.4).

## 5.3 Creating Buy-in

One of the major objectives of an assessment is to create "buy-in" for SPI within the organization. As part of the SPICE Trials survey described above (Section 5.2.2), the respondents were requested to give their perceptions about the extent of buy-in for SPI for before the assessment and since the assessment. This was done for four different roles in the organization: participants in the assessment, the organization's technical staff, the organization's management, and the assessment sponsor. A Wilcoxon matched pairs test was used to compare the "before"and "since the assessment"responses [152]. Hypothesis testing was all one-tailed since we expect the effect to be directional: an assessment increases buy-in.

---

[24] Although, it should be noted that the SPICE Trials study was a relatively small scale survey, and therefore its results are only tentative.

The results are summarized in Figure 6. It can be seen that participants in the assessment and the organization's technical staff increased their buy-in since the assessment. However, the organization's management and sponsor were not perceived to have increased their buy-in. This is evidenced by the results in the previous analysis, whereby the insufficient resources and priority were given to the SPI initiatives.

| Role | p-value |
|---|---|
| Participants in the assessment | **0.00** |
| Organization's technical staff | **0.00** |
| Organization's management | 0.18 |
| Assessment sponsor | 0.12 |

**Figure 6:** The ability of the assessment to create "buy-in"for various roles in the organization. A Wilcoxon matched pairs test was used to compare the "before"and "since the assessment"responses [152]. Hypothesis testing was all one-tailed since we expect the effect to be directional: an assessment increases buy-in. Statistically significant results are bolded in the p-value column.

In summary then, the assessments did not manage to significantly increase the buy-in of the organization's management nor sponsor. This may explain the low priority and low resources given to the SPI effort in many of the assessments, but may also be due to the management and sponsor already having "bought-in"before the assessment, and hence their support of the assessment already existed (and therefore the assessment did not lead to significant changes in that). More encouraging is that the assessment increased buy-in among the participants in the assessment and the technical staff.

The results from the SEI survey (see Section 5.2.1) show that there were increases in buy-in among all four roles. What is clear though is that support for SPI increased most markedly among the people who actually participated in the assessments.

## 5.4 Factors Affecting Improvement Success

In the CMM study (Section 5.2.1), respondents were asked about the extent of success their organizations have had in addressing the findings and recommendations that were raised as a result of the assessment. A series of questions were also asked about factors that may have an impact on SPI Success. These questions can be categorized as *Organizational Factors* or *Barriers*. Subsequently, bivariate associations between each of these factors and SPI Success were investigated[25]. Only the results of statistical testing of this association were presented. A chi-square test was used for these purposes[26]. The "Organizational Factors"and "Barriers"that were found to be statistically significant are summarized in Figure 7.

---

[25] A bivariate association looks at the relationship between only two variables.

[26] A chi-square test, as applied in the SEI study, is a statistical test that is used to find out whether there is an association between two variables when both variables are on a categorical scale. In the bivariate analysis that was performed, one variable was either an *Organizational Factor* or a *Barrier*, and the second was *SPI Success*. They are categorical in the sense that they were responses to a question with a fixed number of response options (categories).

| Organizational Factors |
|---|
| Senior Management Monitoring of SPI |
| Compensated SPI Responsibilities |
| SPI Goals Well Understood |
| Technical Staff Involved in SPI |
| SPI People Well Respected |
| Staff Time/Resources Dedicated to Process Improvement |

| Barriers |
|---|
| Discouragement About SPI Prospects |
| SPI Gets in the Way of "Real" Work |
| "Turf Guarding" Inhibits SPI |
| Existence of Organizational Politics |
| Assessment Recommendations Too Ambitious |
| Need Guidance About How to Improve |
| Need More Mentoring and Assistance |

**Figure 7:** Organizational factors and barriers that were found to be related to SPI Success in [83].

Another study was performed as a follow up to organizations that performed assessments using ISO/IEC 15504 [65] (see Section 5.2.2). This found that the more an organization's SPI effort is determined by the findings of an assessment, the greater the extent to which the assessment findings are successfully addressed during the SPI effort. Therefore, it is important to ensure that the SPI effort is determined by the assessment findings.

To increase the possibility that the assessment's findings determine the SPI effort of the organization, the following factors were found to be important:

- Senior management monitoring of SPI
- Compensated SPI responsibilities
- Ensuring that SPI goals are well understood
- Technical staff involvement in SPI
- Staff and time resources should be made available for SPI
- SPI people well respected

Stelzer and Mellis [160] provide a comprehensive review of the literature with the objective of identifying the factors affecting the success of organizational change in the context of SPI. They came up with ten factors that were perceived to be important, and these are summarized in Table 3. They then carefully reviewed 56 published case reports of SPI using ISO 9000 and the SW-CMM to produce a ranking of the ten factors. The ranks are based on the percentage of case reports that mention a particular factor as important. The rankings are summarized in Table 4.

| **Success Factor of Organizational Change** | **Explanation** |
|---|---|
| Change agents and opinion leaders | Change agents initiate and support the improvement projects at the corporate level, opinion leaders at a local level. |
| Encouraging communication and collaboration | Degree to which communication efforts precede and accompany the improvement program (communication) and degree to which staff members from different teams and departments cooperate (collaboration). |
| Management commitment and support | Degree to which management at all organizational levels sponsor the change. |
| Managing the improvement project | Degree to which a process improvement initiative is effectively planned and controlled. |
| Providing enhanced understanding | Degree to which knowledge of current software processes and interrelated business activities is acquired and transferred throughout the organization. |
| Setting relevant and realistic objectives | Degree to which the improvement efforts attempt to contribute to the success of the organization (relevant) and degree to which the objectives may be achieved in the foreseeable future (realistic). |
| Stabilizing changed processes | Degree to which software processes are continually supported, maintained, and improved at a local level. |
| Staff involvement | Degree to which staff members participate in the improvement activities. |
| Tailoring improvement initiatives | Degree to which improvement efforts are adapted to the specific strengths and weaknesses of different teams and departments. |
| Unfreezing the organization | Degree to which the "inner resistance" of an organizational system to change is overcome. |

**Table 3:** Factors affecting the success of organizational change in SPI (from [160]).

| Success Factor | ISO cases (n = 25) | | CMM cases (n = 31) | | All cases (n = 56) | |
|---|---|---|---|---|---|---|
| | Percentage | Rank | Percentage | Rank | Percentage | Rank |
| Management commitment and support | 84% | 1 | 97% | 1 | 91% | 1 |
| Staff involvement | 84% | 1 | 84% | 8 | 84% | 2 |
| Providing enhanced understanding | 72% | 3 | 87% | 6 | 80% | 3 |
| Tailoring improvement initiatives | 68% | 4 | 90% | 3 | 80% | 3 |
| Managing the improvement project | 56% | 6 | 94% | 2 | 77% | 5 |
| Change agents and opinion leaders | 52% | 7 | 90% | 3 | 73% | 6 |
| Stabilizing changed processes | 52% | 7 | 90% | 3 | 73% | 6 |
| Encouraging communication and collaboration | 64% | 5 | 74% | 9 | 70% | 8 |
| Setting relevant and realistic objectives | 44% | 9 | 87% | 6 | 68% | 9 |
| Unfreezing the organization | 24% | 10 | 52% | 10 | 39% | 10 |

**Table 4:** Ranking of success factors (from [160]).

As can be seen across the three studies there is considerable consistency in the factors that have been found to be important in ensuring the success of an SPI initiative. We can distill the following factors that are the most critical:

- Management commitment and support of SPI (as a demonstration of commitment this would include management monitoring of the initiative and making resources available).
- Involvement of technical staff in the SPI effort.
- Ensuring that staff understand the current software processes and their relationship to other business activities.
- Clear SPI goals that are understood by the staff.
- Tailoring the improvement initiatives.
- Respected SPI staff (change agents and opinion leaders)

## 5.5 The Cost of Software Process Improvement

The most common and interpretable measures of the costs of SPI are in terms of dollars and/or effort. A recent study sponsored by the US Air Force [16] found that government organizations tend to characterize investments in process improvement in terms of costs, whereas industry tend to characterize it in terms of effort expended on SPI activities. In some cases, cost measures such as calendar months have also been used. The studies that we summarize below show the costs of SPI using different approaches. The amount of detail that we can present is directly a function of the amount of publicly available information.

| Ref. | Organization & SPI Program | Costs |
|---|---|---|
| [102] | • SPI effort at the Software Engineering Division of Hughes Aircraft<br>• The division had 500 professional employees at the time | • The assessment itself cost US$45,000<br>• Cost of a 2 year SPI program was US$400,000<br>• Implementation of the action plan to move from ML1 to ML2 was 18 calendar months |
| [171] | • SPI effort led by the Schlumberger Laboratory for Computer Science | Large engineering centers (120-180 engineers) have 1-5 full-time staff on SPI<br>Smaller centers (50-120 engineers) have up to 3 full-time staff on SPI |
| [16]<br>[17] | • Data was collected from 33 companies using questionnaires and/or interviews. | The authors present examples of data on the costs of activities related to SPI.<br>For example, some organizations increased from 7% to 8% of total effort on data collection, and increase upto 2% of project costs on fixing design defects. |
| [18] | Corporate-wide SPI effort at AlliedSignal Aerospace starting in 1992 | Using data on SEPG investment measured in full-time equivalent headcount for 8 sites, the maximum was 4% |
| [41] | • Organization is the Software Systems Laboratory in Raytheon, employing 400 software engineers<br>• SPI initiative started in 1988; results reported after five years<br>• Organization has progressed from Level 1 to Level 3 during that period | US$1 million invested every year |

**Figure 8:** Organizational experiences illustrating the costs of SPI.

### 5.5.1 Costs of Improvement Based on the CMM

A number of companies have publicized the cost details of their process improvement efforts based on the CMM. Some of these are summarized in Figure 8. Another study conducted at the SEI determined the amount of time it takes organizations to increase their maturity levels on the CMM for the first three levels [91]. The distribution of assessments that used the original SPA method and the replacement CBA IPI method in the data set is not clear however, and whether any differences in method would have had any effect on the time it takes to move up one maturity level.

Two groups of organizations were identified: those that moved from level 1 to level 2, and those that moved from level 2 to level 3. On average, it takes organizations 30 months to move from level 1 to level 2. Those organizations, however, varied quite dramatically in the amount of time it takes to move up one maturity level. A more outlier resistant measure would be the median. In this case, the median was 25 months. Organizations that moved from level 2 to level 3 took on average 25 months (the median was also 25 months).

It is plausible that the size of the organization would have an impact on the number of months it takes to move from one maturity level to another. The variation in the size of the organizational units that were assessed was not given in the report however. Therefore, these results should be taken as general guidelines to check an organization's own estimates of the time it takes to move up the maturity ladder.

Another study of US companies found results that are consistent with those mentioned above [16]. It was found that organizations at level 2 spend between 12 to 36 months at level 1with an average of 21 months, and organizations at level 3 had spent 22-24 months at level 1 with an average of 23 months. Organizations at level 3 spent from 12 to 20 months at level 2 with an average of 17.5 months. This is corroborated with data from the improvement efforts at AlliedSignal [18] where advancement from Level 1 to 2 and from Level 2 to Level 3 took 12-14 months across different sites.

### 5.5.2 Costs of Registration to ISO 9001

A multiple regression model has recently been constructed to estimate the effort it would take an organization to meet the requirements of ISO 9001 [145]. Data was collected from 28 software organizations that were registered to ISO 9001 in Canada and the USA. There are two inputs to the model: (a) the size of the organization in number of employees, and (b) the degree of non-compliance to ISO 9001 clauses. Both sets of data were collected by questionnaire and a sample of responses were verified with the respondents to increase confidence in the reliability of the responses. The model to predict effort in man-months is:

$$\text{Ln (effort)} = -2.793 + 0.692 * \text{Ln}(x_1) + 0.74 * \text{Ln}(x_2)$$

where:

$x_1$ = number of employees within the scope of registration

$x_2$ = degree of compliance of the organization to the ISO 9001 clauses prior to the improvement effort

The model was validated using data collected from five organizations that were not included in the model development sample. A brief comparison of the model prediction versus the actual effort is given in Figure 9.

| Org. # | Size | Non-compliance (%) | Predicted | Actual | Residual |
|---|---|---|---|---|---|
| 1 | 175 | 35% | 30.3 | 31.2 | 0.9 |
| 2 | 108 | 15% | 11.6 | 13 | 1.4 |
| 3 | 170 | 30% | 26.5 | 27 | 0.5 |
| 4 | 45 | 100% | 25.8 | 36 | 10.2 |
| 5 | 100 | 70% | 34.4 | 37 | 2.6 |

**Figure 9:** Comparison of actual versus predicted effort for ISO 9001 registration.

### 5.5.3 Other Models

Jones [108] presents the stages of Software Productivity Research (SPR) Inc's improvement model as follows:

- **Stage 0:** Software Process Assessment and Baseline
- **Stage 1:** Focus on Management Technologies
- **Stage 2:** Focus on Software Processes and Methodologies
- **Stage 3:** Focus on New Tools and Approaches
- **Stage 4:** Focus on Infrastructure and Specialization
- **Stage 5:** Focus on Reusability
- **Stage 6:** Focus on Industry Leadership

From the data that SPR has collected, the costs per capita (in US $) for small and large organizations to progress through the stages are presented in Table 5, and the time it takes to make these progressions are summarized in Table 6.

| STAGE | SMALL<br>< 100<br>STAFF | MEDIUM<br>101 -- 1000<br>STAFF | LARGE<br>1001 – 10000<br>STAFF | GIANT<br>>10000<br>STAFF |
|---|---|---|---|---|
| Stage 0 -- Assessment/baseline | $100 | $125 | $150 | $200 |
| Stage 1 -- Management | $1,500 | $5,000 | $5,000 | $8,000 |
| Stage 2 -- Methods/Process | $1,500 | $2,500 | $3,000 | $3,500 |
| Stage 3 -- New Tools | $5,000 | $7,500 | $15,000 | $25,000 |
| Stage 4 -- Infrastructure | $1,000 | $1,500 | $3,500 | $5,000 |
| Stage 5 -- Reusability | $500 | $3,000 | $6,000 | $7,500 |
| Stage 6 -- Industry Leadership | $1,500 | $2,500 | $3,000 | $4,000 |
| Approximate total | $11,100 | $22,125 | $35,650 | $48,200 |

**Table 5:** Costs per capita to progress through SPR's stage model. These include training, consulting fees, capital equipment, software licenses, and improvements in office conditions. This table is presented in [108].

| STAGE | SMALL<br>< 100<br>STAFF | MEDIUM<br>101 -- 1000<br>STAFF | LARGE<br>1001 -- 10000<br>STAFF | GIANT<br>>10000<br>STAFF |
|---|---|---|---|---|
| Stage 0 -- Assessment/baseline | 2 | 2 | 3 | 4 |
| Stage 1 -- Management | 3 | 6 | 9 | 12 |
| Stage 2 -- Methods/Process | 4 | 6 | 9 | 15 |
| Stage 3 -- New Tools | 4 | 6 | 9 | 12 |
| Stage 4 -- Infrastructure | 3 | 4 | 6 | 9 |
| Stage 5 -- Reusability | 4 | 6 | 12 | 12 |
| Stage 6 -- Leadership | 6 | 8 | 9 | 12 |
| TOTAL | 26 | 38 | 57 | 76 |

**Table 6:** Calendar months it takes to progress through the SPR model's stages. This table is presented in [108].

## 5.6 Summary

The following are the main points from the review of the literature on SPI:

- There is no evidence supporting or refuting that the ordering of processes or process capabilities in contemporary best practice models really reflects the natural evolution of successful software organizations. However, this does not necessarily diminish the utility of the "stage" concept of best practice models: assessed organizations find the stages concept to be very useful for prioritizing their improvement efforts.
- Approximately one third of respondents on SPI surveys report marked changes in their organizations as a consequence of their assessments.
- Organizations seem to have (unrealistically) high expectations from assessment-based SPI, and these need to be managed better.
- Despite the above, sponsors of assessments and assessment participants still remain enthusiastic about SPAs.
- Assessments increase 'buy-in' for SPI for the participants in the assessments and the organization's technical staff. This is certainly encouraging as these are the individuals who are most likely to be skeptical about SPI and resist change to yet another management initiative.

- The most important factors that affect the success of SPI initiatives are (there are other factors that are important, but the ones below seem to be the most important):
    - Management commitment and support of SPI (as a demonstration of commitment this would include management monitoring of the initiative and making resources available).
    - Involvement of technical staff in the SPI effort.
    - Ensuring that staff understand the current software processes and their relationship to other business activities.
    - Clear SPI goals that are understood by the staff.
    - Tailoring the improvement initiatives.
    - Respected SPI staff (change agents and opinion leaders)
- We have provided some general guidelines on the costs of SPI that may be useful in managing the expectations of organizations better.

# 6 The Dimensions of Process Capability

In Section 4.3 we saw that the measurement properties of software process capability measures represents an important criterion for evaluating SPAs. The embedded assumption in contemporary best practice models is that process capability is a unidimensional construct. This can be seen in the unidimensional nature of the SW-CMM and the capability scale of ISO/IEC 15504. In this section we review studies that test this assumption.

Studies that investigate dimensionality have been exploratory in nature. For this, principal components analysis (PCA) has been used as the investigative technique [111]. PCA identifies groups in a set of variables.

One of the first such studies was performed by Curtis [35]. He collected data from a large number of individuals in three organizations about their software engineering practices following the SW-CMM. Questions on satisfying the goals of the KPAs were posed. A PCA was performed on each organization's data individually, and then the overall results were collectively interpreted. His results are shown in Figure 10.

One immediate conclusion is that process capability, as defined in the SW-CMM, is a multidimensional construct. Also, it is clear that the goals of individual KPAs do not load on the same component. For example, the "planfulness" dimension is covered by goals from three separate KPAs, the "Coordinated Commitments" dimension is covered by goals from two different KPAs, and the "Process Definition" dimension by goals from three different KPAs. Although, there is also some consistency. The "Subcontractor Management", "Quality Assurance", and "Configuration Management" dimensions are composed of goals from the same KPA.

Clark, in his PhD thesis [25], provided the correlation matrix for 13 SW-CMM KPAs for data collected from 50 organizations. This can be used to perform a PCA, which we did. The results are shown inFigure 11.

It is clear that the dimensionality of SW-CMM process capability is different between these two studies. For example, Clark's data indicate that the Intergroup Coordination and Software Project Tracking are different dimensions, whereas Curtis' results shows that they are strongly related at the goal level. Apart from the Subcontractor Management KPA and the anomaly above, one would say that the KPAs studied by Curtis load on the same dimension in Clark's data. For example, from Clark's data, configuration management, quality assurance, and training are all in the same dimension, whereas they represent three different dimensions in Curtis' results.

|  |  |  |
|---|---|---|
| **Planfulness** |  |  |
| Project Planning | G1 | Software estimates are documented for each use in planning and tracking the software project. |
| Project Planning | G2 | Software project activities and commitments are planned and documented. |
| Project Tracking | G1 | Actual results and performance are tracked against the software plans. |
| Project Tracking | G2 | Corrective actions are taken and managed to closure when actual results and performance deviate significantly from the software plans. |
| Integrated Software Management | G2 | The project is planned and managed according to the project's defined software process. |
| **Coordinated Commitments** |  |  |
| Project Planning | G3 | Affected groups and individuals agree to their commitments related to the software project. |
| Project Tracking | G3 | Changes to software commitments are agreed to by the affected groups and individuals. |
| Intergroup Coordination | G1 | The customer's requirements are agreed to by all affected groups. |
| Intergroup Coordination | G2 | The commitments between the engineering groups are agreed to by the affected groups. |
| Intergroup Coordination | G3 | The engineering groups identify, track, and resolve intergroup issues. |
| **Subcontractor Management** |  |  |
| Subcontractor Management | G1 | The prime contractor selects qualified software subcontractors. |
| Subcontractor Management | G2 | The prime contractor and the software subcontractor agree to their commitments to each other. |
| Subcontractor Management | G3 | The prime contractor and the software subcontractor maintain ongoing communications. |
| **Quality Assurance** |  |  |
| Software Quality Assurance | G1 | Software quality assurance activities are planned. |
| Software Quality Assurance | G2 | Adherence of software products and activities to the applicable standards, procedures, and requirements is verified objectively. |
| Software Quality Assurance | G3 | Affected groups and individuals are informed of software quality assurance activities and results. |
| Software Quality Assurance | G4 | Noncompliance issues that cannot be resolved within the software project are addressed by senior management. |
| **Configuration Management** |  |  |
| Configuration Management | G1 | Software configuration management activities are planned. |
| Configuration Management | G2 | Selected software work products are identified, controlled, and available. |
| Configuration Management | G3 | Changes to identified software work products are controlled. |
| Configuration Management | G4 | Affected groups and individuals are informed of the status and content of software baselines. |
| **Process Definition** |  |  |
| Organization Process Focus | G1 | Software process development and improvement activities are coordinated across the organization. |
| Organization Process Focus | G2 | The strengths and weaknesses of the software processes used are identified relative to a process standard. |
| Organization Process Focus | G3 | Organization-level process development and improvement activities are planned. |
| Organization Process Definition | G1 | A standard software process for the organization is developed and maintained. |

|  |  |  |
|---|---|---|
| Organization Process Definition | G2 | Information related to the use of the organization's standard software process by the software projects is collected, reviewed, and made available. |
| Integrated Software Management | G1 | The project's defined software process is a tailored version of the organization's standard software process. |
| **Training** |  |  |
| Training Program | G1 | Training activities are planned. |
| Training Program | G2 | Training for developing the skills and knowledge needed to perform software management and technical roles is provided. |
| Training Program | G3 | Individuals in the software engineering group and software-related groups receive the training necessary to perform their roles. |

**Figure 10:** Results of the Curtis PCA. The G's are references to the goals of the associated KPA.

|  | PC 1 | PC 2 | PC 3 |
|---|---|---|---|
| Requirements Management | .224205 | **.793343** | .167607 |
| Software Project Planning | **.661053** | .349884 | .443339 |
| Software Project Tracking and Oversight | **.705950** | .420637 | .414642 |
| Software Subcontract Management | .056427 | -.056777 | **-.917932** |
| Software Quality Assurance | **.696583** | .460522 | .175482 |
| Software Configuration Management | **.521998** | .493161 | -.049939 |
| Organization Process Focus | **.910695** | .153472 | -.070893 |
| Organization Process Definition | **.924298** | .047921 | -.013578 |
| Training Program | **.605227** | .337231 | .216687 |
| Integrated Software Management | **.841650** | .148948 | -.058104 |
| Software Product Engineering | .122982 | **.747144** | .012944 |
| Intergroup Coordination | .230563 | **.719089** | .072065 |
| Peer Reviews | **.900477** | .213542 | .027097 |
| Expl.Var | 5.384027 | 2.669933 | 1.334092 |
| Prp.Totl | .414156 | .205379 | .102622 |

**Figure 11:** Factor structure for the first 13 KPAs from Clark's thesis.

There are a number of ways of explaining these results, unfortunately none of them substantive. First, Curtis collected data from individuals from the same organization. Therefore, all the respondents within a PCA were responding about the same practices (i.e., the unit of observation was the individual rather than the organization). This type of design is useful for reliability studies, but ideally the unit of observation should be the organization or the project. Such was the case in Clark's thesis, and therefore one is tempted to conclude that Clark's results are more likely to be stable. In addition, the two studies did use different measurement approaches. Curtis asked questions on the satisfaction of KPA goals. Clark asked about the frequency of implementation of the KPAs.

Another study by El Emam investigated the dimensionality of the ISO/IEC 15504 capability scale [60]. He found that to be two dimensional, with the attributes in the first three levels of the capability scale representing one dimension, and the attributes in levels 4 and 5 representing the second dimension.

It is clear from the above studies that process capability is not a unidimensional construct. Beyond that it is quite difficult to draw strong conclusions about the dimensionality of the SW-CMM practices. For ISO/IEC 15504 two dimensions have been identified. If indeed further research confirms and identifies the dimensions of process capability, then it may be plausible to define different measures of capability with higher reliability and validity rather than attempt to have single universal measures.

# 7 The Reliability of Process Capability Measures[27]

Reliability is an enduring concern for software process assessments. The investment of time, money, and personal effort needed for assessments and successful software process improvement is quite non-trivial, and decisions based on assessment results are often far-reaching. Organizations and acquirers of software systems must be confident that the assessment results are well-founded and repeatable.

Reliability is defined as the extent to which the same measurement procedure will yield the same results on repeated trials and is concerned with random measurement error [26]. This means that if one were to repeat the measurement under similar or compatible conditions the same outcomes would emerge.

There has been a concern with the reliability of assessments. For example, Card discusses the reliability of Software Capability Evaluations in an earlier article [24], where he commented on the inconsistencies of the results obtained from assessments of the same organization by different teams. At the same time, another report noted that comparisons of SCE outcomes being cited as evidence of their lack of reliability were frequently based on SCEs performed on different subsets of the organization (or at least had different interviewees), and therefore one would not expect identical results [159]. Mention is also made of reliability in a contract award situation where emphasis is placed on having one team assess different contractors to ensure consistency [148]. Bollinger and McGowan [13] criticize the extent to which the scoring scheme used in the SEI's Software Capability Evaluations contributes towards reduced reliability (see also [103] for a response). The Interim Profile method of the SEI [170] includes specific indicators to evaluate reliability. Furthermore, a deep concern with reliability is reflected in the empirical trials of the prospective SPICE standard whereby the evaluation of the reliability of SPICE-conformant assessments is an important focus of study [52].

One important implication of the extent of unreliability is that the ratings obtained from an assessment is only one of the many possible ratings that would be obtained had the organization been repeatedly assessed. This means that, for a given level of confidence that one is willing to tolerate, an assessment rating has a specific probability of falling within a range of ratings. The size of this range increases as reliability decreases. This is illustrated in Figure 12.

---

[27] This section is based partially on material from [52][85][68][59][79].

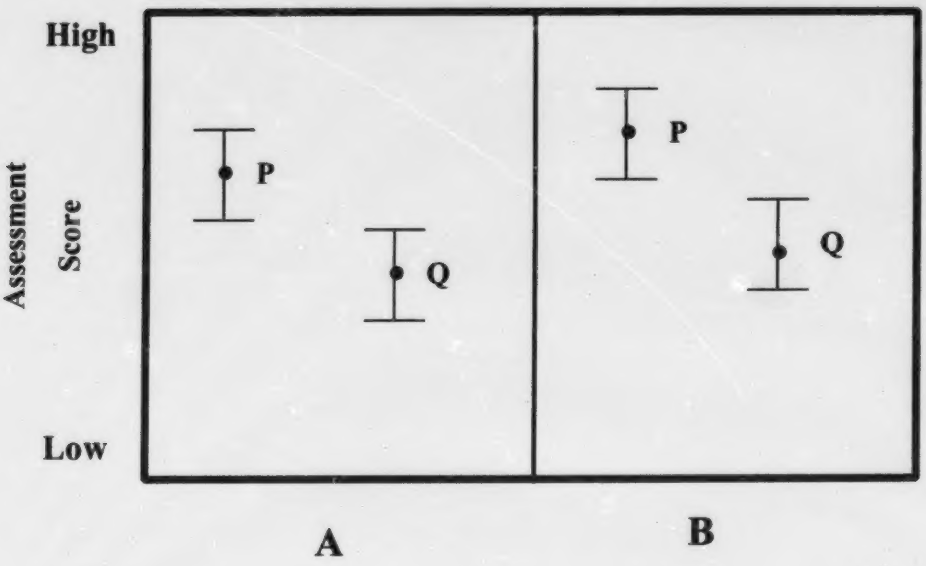


**Figure 12:** Example hypothetical assessment ratings with confidence intervals.

Assume Figure 12 shows the profiles of two organizations, A and B, and that P and Q are two different processes being assessed. Due to random measurement error, the ratings obtained for each process are only one of the many possible ratings that would be obtained had the organization been repeatedly assessed. While obtained ratrings for organization B are in general higher than those of organization A, this may be an artifact of chance. Without consideration of random measurement error, organization A may be unfairly penalized in a contract award situation. Turning to a self-improvement scenario, assume that Figure 12 shows the profiles of one organization at two points in time, A and B. At time A, it may seem that the rating for process Q is much lower than for process P. Thus, the organization would be tempted to pour resources on improvements in process Q. However, without consideration of random measurement error, one cannot have high confidence about the extent to which the difference between P and Q ratings is an artifact of chance. Furthermore, at time B, it may seem that the organization has improved. However, without consideration of random measurement error, one cannot have high confidence about the extent to which the difference between A and B ratings (for processes P and Q) are artifacts of chance.

The above examples highlight the importance of evaluating the extent of reliability of software process assessments. A recent study also found evidence that more reliable assessments are indeed less costly [61]. The reason for that is when assessments are reliable (i.e., the assessors agree in their initial ratings), the consolidation phase progresses faster with less consensus building, hence resulting in an overall reduction in cost.[28]

## 7.1 Reliability Theory and Methods

Two theoretical frameworks for ascertaining the extent of reliability are presented below: (a) the classical test theory framework, and (b) the generalizability theory framework. Associated with each theoretical framework are a number of empirical research methods that can be applied.

---

[28] Note that this effect was identified for certain types of processes rather than for all processes that were studied.

### 7.1.1 Classical Test Theory Framework

Classical test theory states that an observed rating consists of two additive components, a true rating and an error: $X = T + E$. Thus, X would be the rating obtained in an assessment, T is the mean of the theoretical distribution of X ratings that would be found in repeated assessments of the same organization[29], and E is the error component. The reliability of measurement is defined as the ratio of true variance to observed variance.

| | | **Number of Assessment Procedures Required** | |
|---|---|---|---|
| | | **One** | **Two** |
| **Number of Assessment Occasions Required** | **One** | Split-Halves<br>Internal Consistency | Alternative Forms (immediate) |
| | **Two** | Test-retest | Alternative Forms (delayed) |

**Figure 13:** A classification of classical reliability estimation methods.

There are four methods for estimating reliability under this framework. All of the four methods attempt to determine the proportion of variance in a measurement scale that is systematic. The different methods can be classified by the number of different assessment procedures necessary and the number of different assessment occasions necessary to make the estimate, as summarized in Figure 13. These methods are briefly described below:

1. Test-Retest Method

   This is the simplest method for estimating reliability. In an assessment one would have to assess each organization's capability at two points in time using the same assessment procedure (i.e., the same instrument, the same assessors, and the same assessment method). Reliability would be estimated by the correlation between the ratings obtained on the two assessments. Test-retest reliability suffers from three disadvantages. First is the expense of conducting assessments at more than point in time. Given that the cost of assessments is an enduring concern [12][106], the costs of repeated assessments for the purpose of estimating reliability would generally be unacceptable. Second, it is not obvious that a low reliability coefficient obtained using test-retest really indicates low reliability. For example, a likely explanation for a low coefficient is that the organization's software process capability has changed between the two assessment occasions. For instance, the initial assessment and results may (should) sensitize the organization to specific weaknesses and prompt them to initiate an improvement effort that influences the results of the subsequent assessment. Finally, carry-over effects between assessments may lead to an over-estimate of reliability. For instance, the reliability coefficient can be artificially inflated due to memory effects. Examples of memory effects are the assessees knowing the 'right' answers that they have learned from the previous assessments, and assessors remembering responses from previous assessments and, deliberately or otherwise, repeating them in an attempt to maintain consistency of results.

2. Alternative Forms Method

   Instead of using the same assessment procedure on two occasions, the alternative forms method stipulates that two alternative assessment procedures be used. This can be achieved, for example, by using two different assessment instruments or having two alternative, but equally qualified, assessment teams. This method can be characterized either as immediate (where the two assessments are concurrent in time), or delayed (where the two assessments are separated in time). The correlation coefficient (or some other measure of association) is then used as an estimate of reliability of *either* of the alternative forms.

---

[29] In practice the true score can never be really known since it is generally not possible to obtain a large number of repeated assessments of the same organization. If one is willing to make some assumptions (e.g., an assumption of linearity), however, point estimates of true scores can be computed from observed scores [125].

3. Split-Halves Method

   With the split-halves method, the total number of items in an assessment instrument are divided into two halves and the half-instruments are correlated to get an estimate of reliability. The halves can be considered as approximations to alternative forms. A correction must be applied to the correlation coefficient though, since that coefficient gives the reliability of each half only. One such correction is known as the Spearman-Brown prophecy formula [136]. A difficulty with the split-halves method is that the reliability estimate depends on the way the instrument is divided into halves. For example, for a 10 question instrument there are 126 possible different splits, and hence 126 different split-halves reliability estimates. The most common procedure is to take even numbered items on an instrument as one part and odd numbered ones as the second part.

4. Internal Consistency Methods

   With methods falling under this heading, one examines the covariance among all the items in an assessment instrument. By far the most commonly used internal consistency estimate is the Cronbach alpha coefficient [31].

Since there exists more than one classical methods for estimating reliability, a relevant question is "which method(s) should be used ?" One way to answer this is to consider what the research community perceives to be the most important reliability evaluation techniques. If we take the field of Management Information Systems (MIS) as a reference discipline (in the sense that MIS researchers are also concerned with software processes, their measurement, and their improvement), then some general statements can be made about the perceived relative importance of the different methods.

In MIS, researchers developing instruments for measuring software processes and their outcomes tend to report the Cronbach alpha coefficient most frequently [168]. Furthermore, some researchers consider the Cronbach alpha coefficient to be the most important [151].

Examples of instruments with reported Cronbach alpha coefficients are those for measuring user information satisfaction [105][162], user involvement [4][6], and perceived ease of use and usefulness of software [38]. Moreover, recently, reliability estimates using other methods have also been reported, for example, test-retest reliability for a user information satisfaction instrument [80], and for a user involvement instrument [164].

In software engineering, the few studies that consider reliability report the Cronbach alpha coefficient. For example, the reliability estimate for a requirements engineering success instrument [55], for an organizational maturity instrument [53], and for level 2 and 3 of the SEI maturity questionnaire [103].

### 7.1.2 Generalizability Theory Framework

The different classical methods for estimating reliability presented above vary in the factors that they subsume under error variance. Some common sources of random measurement are presented in Figure 14 [52]. This means that the use of different classical methods will yield different estimates of reliability.

Generalizability theory [32], however, allows one to *explicitly consider multiple sources of error simultaneously and estimate their relative contributions*. In the context of process assessments, the theory would be concerned with the accuracy of generalizing from an organization's obtained rating on an assessment to the average rating that the organization would have received under all possible conditions of assessment (e.g., using different instruments, different assessment teams, different team sizes). This average rating is referred to as the *universe rating*. All possible conditions of assessment are referred to as the *universe of assessments*. A set of measurement conditions is called a *facet*. Facets relevant to assessments include the assessment instrument used, the assessment team, and assessment team size.

Generalizability theory uses the factorial analysis of variance (ANOVA) [132] to partition an organization's assessment rating into an effect for the universe rating, an effect for each facet or source of error, an effect for each of their combinations, and other "random" error. This can be contrasted to simple ANOVA, which is more analogous to the classical test theory framework. With simple ANOVA the variance is partitioned into "between" and "within". The former is though of as systematic variance or signal. The latter is thought of as random error or noise. In the classical test theory framework one similarly partitions the total variance into true rating and error rating.

Suppose, for the purpose of illustration, one facet is considered, namely assessment instrument. Further, suppose that in an evaluation study two instruments are used and N organizations are assessed using each of the two instruments. In this case, one intends to generalize from the two instruments to all possible instruments. The results of this study would be analyzed as a two-way ANOVA with one observation per cell (e.g., see [132]). The above example could be extended to have multiple facets (i.e., account for multiple sources of error such as instruments and assessors).

| Source of Error | Description |
|---|---|
| **Different Occasions** | Assessment ratings may differ across time. Instability of assessment ratings may be due to temporary circumstances and/or actual process changes. |
| **Different Assessors** | Assessment ratings may differ across assessors (or assessment teams). Lack of repeatability of assessment ratings may be due to the subjectivity in the evaluations and judgment of particular assessors (i.e., do different assessors make the same judgments about an organization's processes?). |
| **Different Instrument Contents** | Assessment ratings may differ across instruments. Lack of equivalence of instruments may be due to the questions in different instruments not being constructed according to the same content specifications (i.e., do different instruments have questions that cover the same content domain?). |
| **Within Instrument Contents** | Responses to different questions or subsets of questions within the same instrument may differ among themselves. One reason for these differences is that questions or subsets of questions may not have been constructed to the same or to consistent content specifications.<br>Regardless of their content, questions may be formulated poorly, may be difficult to understand, may not be interpreted consistently, etc. |

**Figure 14**: Definition of some sources of error in process assessments.

### 7.1.3 Applications

The studies that have been performed thus far on the reliability of assessments only utilize classical methods. Specifically, the evaluation of internal consistency using the Cronbach alpha coefficient and the alternative forms (immediate) methods. For the latter the alternative form is a different assessor or assessment team. This is commonly called an interrater agreement study where instead of using the correlation coefficient, other approaches are used to quantify reliability.

## 7.2 Internal Consistency

A basic concept for comprehending the reliability of measurement is that of a *construct*. A construct refers to a meaningful conceptual object. A construct is neither directly measurable nor observable. However, the quantity or value of a construct is presumed to cause a set of observations to take on a certain value.

An observation can be considered as a question in a maturity questionnaire (this is also referred to as an *item*). Thus, the construct can be indirectly measured by considering the values of those items.

For example, organizational maturity is a construct. Thus, the value of an item measuring *"the extent to which projects follow a written organizational policy for managing system requirements allocated to software"* is presumed to be caused by the true value of organizational maturity. Also, the value of an item measuring *"the extent to which projects follow a written organizational policy for planning software projects"* is presumed to be caused by the true value of organizational maturity. Such a relationship is depicted in the path diagram in Figure 15. Since organizational maturity is not directly measurable, the above two items are intended to estimate the actual magnitude or true rating of organizational maturity.

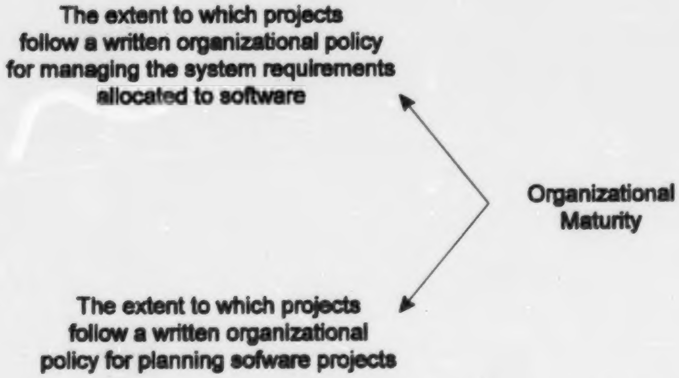

The extent to which projects follow a written organizational policy for managing the system requirements allocated to software
Organizational Maturity
The extent to which projects follow a written organizational policy for planning sofware projects

**Figure 15:** Path diagram depicting the organizational maturity construct and example items for its measurement.

The type of scale used in the most common assessment instruments is a summative one. This means that the individual ratings for each item are summed up to produce an overall rating. One property of the covariance matrix for a summative scale that is important for the following formulation is that the sum of all the elements in the matrix give exactly the variance of the scale as a whole.

One can think of the variability in a set of item ratings as being due to one of two things: (a) actual variation across the organizations in maturity (i.e., true variation in the construct being measured) and this can be considered as the signal component of the variance, and (b) error which can be considered as the noise component of the variance. Computing the Cronbach alpha coefficient involves partitioning the total variance into signal and noise. The proportion of total variation that is signal equals Cronbach alpha.

The signal component of variance is considered to be attributable to a common source, presumably the true rating of the construct underlying the items. When maturity varies across the different organizations, ratings on all the items will vary with it because it is a cause of these ratings. The error terms are the source of unique variation that each item possesses. Whereas all items share variability due to maturity, no two items are considered to share any variation from the same error source.

Unique variation is the sum of the elements in the diagonal of the covariance matrix: $\Sigma\sigma_i^2$ (where i=1..N items). Common variation is the difference between total variation and unique variation: $\sigma_y^2 - \Sigma\sigma_i^2$, where the first term is the variation of the whole scale. Therefore, the proportion of common variance can be expressed as: $(\sigma_y^2 - \Sigma\sigma_i^2)/\sigma_y^2$ . To express this in relative terms, the number of elements in the matrix must be considered. The total number of elements is $N^2$, and the total number of elements that are communal are $N^2 - N$. Thus the corrected equation for coefficient alpha becomes:

$$\alpha = \frac{N}{(N-1)}\left[1 - \sum \sigma_i^2 / \sigma_y^2\right]$$

The Cronbach alpha coefficient varies between 0 and 1. If there is no true rating but only error in the items, then the variance of the sum will be the same as the sum of variances of the individual items. Therefore, coefficient alpha will be equal to zero (i.e., the proportion of true ratings in the scale is zero percent). If all items are perfectly reliable and measure the same thing, then coefficient alpha is equal to one (i.e., the proportion of true rating in the scale is 100 percent).

What a satisfactory level of internal consistency is depends on how a measure is being used. In the early stages of the research on assessment instruments, reliabilities of 0.7 or higher are considered sufficient. For basic research, a value of 0.8 is acceptable. However, in applied settings where important decisions are made with respect to assessment ratings, a reliability of 0.9 is considered a minimum [136].

To illustrate the calculation of Cronbach's alpha through an example, let's assume that we have an instrument with five items. Also, let the variances of each of these items be 4.32, 3.86, 2.62, 2.55, and 2.46. The sum of the variances, $\Sigma\sigma_i^2$ , is 15.81. For this five item instrument, the variance of the sum of the 5 items, $\sigma_y^2$ , was say 53.95. Then the Cronbach alpha coefficient would be 1.25(1-(15.81/53.95)) = 0.88. The values for the sample variances can be computed as $\left(\sum(x_j-\bar{x})^2\right)/\,NOBS-1$, where the $x_j$'s are the actual values for each observation j (j=1 .. NOBS), and $\bar{x}$ is the mean of all observations, and NOBS is the total number of observations.

Cronbach's alpha is a generalization of a coefficient introduced by Kuder and Richardson to estimate the reliability of scales composed of dichotomously scored items. Dichotomous items are scored one or zero depending on whether the respondent does or does not endorse the particular characteristic under investigation. To determine the reliability of scales composed of dichotomously scored items, the Kuder-Richardson formula (symbolized KR20) is [2]:

$$KR20=\frac{N}{(N-1)}\left[1-\sum p_i q_i/\sigma_y^2\right]$$

where $N$ is the number of dichotomous items; $p_i$ is the proportion responding positively to the $i^{th}$ item; $q_i$ is equal to $1-p_i$; and $\sigma_y^2$is equal to the variance of the total composite. Since KR20 is simply a special case of alpha, and it has the same interpretation as alpha.

A series of studies investigated the internal consistency of the CMM maturity questionnaire (the questionnaire described in [101]) [103][79], the initial version of the ISO/IEC 15504 capability dimension [79], and the second version of the ISO/IEC 15504 capability dimension [60].

Before presenting the results, it is worthwhile to note that the Cronbach alpha coefficient depends on the number of items in an instrument. Therefore, to compare different instruments, or present their internal consistency in a comparable manner, the coefficient is adjusted to an instrument of the same size.

Figure 16 shows the internal consistency results for the first set of studies [103][79]. The coefficients are presented for two sized instruments, an 85 item instrument (to be compatible with the maturity questionnaire size), and a 26 item instrument (to be compatible with the ISO/IEC 15504 version 1 capability dimension). In general, it can be seen that these instruments have a remarkably high internal consistency.

| | **1987 maturity questionnaire Data Set 1** | **1987 maturity questionnaire Data Set 2** | **Estimates based on [103]** | **SPICE v1 Capability Dimension Data Set 1** | **SPICE v1 Capability Dimension Data Set 2** |
|---|---|---|---|---|---|
| **85 item instrument** | 0.94 | 0.94 | 0.92 | 0.98 | 0.99 |
| **26 item instrument** | 0.84 | 0.84 | 0.78 | 0.94 | 0.97 |

**Figure 16:** Comparisons of early internal consistency results.

However, as we saw earlier on, the above capability measures are not all unidimensional. The Cronbach alpha coefficient assumes unidimensionality. Therefore, it is not obvious how to interpret the results in Figure 16.

| Attributes up to Level 3 (5 attributes; n=312) | Attributes in Levels 4 and 5 (4 attributes; n=232) |
|---|---|
| 0.89 | 0.90 |

**Table 7:** Cronbach alpha coefficients for different numbers of attributes.

As noted earlier, a subsequent study [60] found that the ISO/IEC 15504 capability dimension actually consisted of two separate dimensions: "Process Implementation" and "Quantitative Process Management". The internal consistency for these are shown in Table 12. Here, the results demonstrate a high internal consistency for each dimension individually.

## 7.3 Interrater Agreement

A number of interrater agreement studies have been performed on software process assessments [57][58][59][62]. All of these were performed in the context of the SPICE Trials.

For conducting interrater agreement studies, the assessment team is divided into two or more groups. Ideally all groups should be equally competent in making attribute ratings. In practice, assessors in each group need only meet minimal competence requirements (described in [104]) since this is more congruent with the manner in which the ISO/IEC 15504 documents would be applied.

Both groups would participate in the preparation of the assessment. During evidence collection, each group would be provided with the same information (e.g., all would be present in the same interviews and provided with the same documentation to inspect) [30], and then they would perform their ratings independently. It is these independent ratings that are used in interrater agreement studies.

Subsequent to the independent ratings, the groups would meet to reach a consensus in their findings and ratings, and produce a consolidated rating which is the final outcome of an assessment. This is the consolidation phase. Subsequently the assessment team may discuss the findings and ratings with the initial interviewees (debriefing), which may lead to refinement of the findings and ratings. Then they present the final results to the organization. Consolidation, debriefing, and reporting are activities that are not necessary for evaluating interrater agreement, but they must be performed so that the organization that sponsored the assessment gets value out of it. Given the expense of an assessment, few, if any, organizations will sponsor assessments that produce no output. General guidelines for conducting interrater agreement studies are given in Table 9.

Since all of the interrater agreement studies that have been conducted thus far were in the context of the SPICE Trials, we briefly review the general rating scheme of ISO/IEC 15504 to aid in understanding the methods used [64]. The rating scheme consists of a 4-point *achievement* scale for each attribute. The four points are designated as F, L, P, N for *Fully Achieved*, *Largely Achieved*, *Partially Achieved*, and *Not Achieved*. A summary of the definition for each of these response categories is given in Table 8.

---

[30] Under this requirement, one group may obtain information that was elicited by the other group, which they would have not asked for. The alternative to this requirement is that the different groups interview the same people at different times to make sure that they only obtain the information that they ask for. However, this requirmeent raises the risk that the interviewees "learn" the right answers to give based on the first interview, or that they volunteer information that was asked by the first group but not the second. Furthermore, from a practical perspective, intervieweing the same people more than once to ask the same questions would substantially increase the cost of assessments, and thus the cost of conducting a study. It is for this reason that these studies are referred to as "interrater" agreement since, strictly speaking, they consider the reliability of ratings, rather than the reliability of whole assessments. The study of "interassessment" agreement would involve accounting for variations in the information that is collected by two (or more) different groups during an assessment.

| **Rating & Designation** | ***Description*** |
|---|---|
| Not Achieved - N | There is no evidence of achievement of the defined attribute. |
| Partially Achieved - P | There is some achievement of the defined attribute. |
| Largely Achieved - L | There is significant achievement of the defined attribute. |
| Fully Achieved - F | There is full achievement of the defined attribute. |

**Table 8:** The four-point attribute rating scale.

Instructions for Conducting An Interrater Agreement Study

- For each process, divide the assessment team into k>=2 groups with at least one person per group.
- The k groups should be selected so that they both meet the minimal assessor competence requirements with respect to training, background, and experience.
- The k groups should use the same evidence (e.g., attend the same interviews, inspect the same documents, etc.), assessment method, and tools.
- Each group examining any physical artifacts should leave them as close as possible (organized/marked/sorted) to the state that the assessees delivered them.
- If evidence is judged to be insufficient, gather more evidence and the k groups should inspect the new evidence before making ratings.
- The k groups independently rate the same process instances.
- After the independent ratings, the k groups then meet to reach consensus and harmonize their ratings for the final ratings profile.
- There should be no discussion among the k groups about rating judgment prior to the independent ratings[31].

**Table 9:** Guidelines for conducting interrater agreement studies.

---

[31] This requirement needs special attention when the assessment method stipulates having multiple consolidation activities throughout an assessment (e.g., at the end of each day in an assessment). Observations that are discussed during such sessions can be judged as organizational strengths or weaknesses, and therefore the ratings of the different groups would no longer be independent. This can be addressed if consolidation is performed independently by the different groups. Then, before the presentation of findings to the organization, overall consolidation of ratings and findings by the different groups is performed.

|  |  | Group 1 |  |  |  |  |
|---|---|---|---|---|---|---|
|  |  | F | L | P | N |  |
| Group 2 | F | $P_{11}$ | $P_{12}$ | $P_{13}$ | $P_{14}$ | $P_{1+}$ |
|  | L | $P_{21}$ | $P_{22}$ | $P_{23}$ | $P_{24}$ | $P_{2+}$ |
|  | P | $P_{31}$ | $P_{32}$ | $P_{33}$ | $P_{34}$ | $P_{3+}$ |
|  | N | $P_{41}$ | $P_{42}$ | $P_{43}$ | $P_{44}$ | $P_{4+}$ |
|  |  | $P_{+1}$ | $P_{+2}$ | $P_{+3}$ | $P_{+4}$ |  |

**Table 10:** 4x4 table for representing *proportions* from an ISO/IEC 15504 interrater agreement study with two groups.

Data from an interrater agreement study of an ISO/IEC 15504 assessment can be represented in a table such as Table 10. Here we have two groups that have independently made a number of ratings on the 4-point scale described above. The table would include the proportion of ratings that fall in each one of the cells.

In this table $P_{ij}$ is the proportion of ratings classified in cell (i,j), $P_{i+}$ is the total proportion for row i, and $P_{+j}$ is the total proportion for column j:

$$P_{i+} = \sum_{j=1}^{4} P_{ij}$$

$$P_{+j} = \sum_{i=1}^{4} P_{ij}$$

The most straightforward approach to evaluating agreement is to consider the proportion of ratings upon which the two groups agree:

$$P_O = \sum_{i=1}^{4} P_{ii}$$

However, this value includes agreement that could have occurred by chance. For example, if the two groups employed completely different criteria for assigning their ratings to the same practices (i.e., if the row variable was independent from the column variable), then a considerable amount of observed agreement would still be expected by chance.

Hartmann [90] notes that percentage (or proportion) agreement tends to produce higher values than other measures of agreement, and discourages its use since the tradition in science is to be conservative rather than liberal. A more detailed analysis of the behavioral literature where proportion agreement was used concluded that large fractions of these observations would be deemed unreliable if corrections for chance were considered [161]. Therefore, in general, the use of percentage or proportion agreement is not recommended as an evaluative measure.

There are different ways for evaluating extent of agreement that is expected by chance. We will consider two alternatives here. The first assumes that chance agreement is due to each of the groups rating randomly at equal rates for each of the categories of the four-point scales. In such a case chance agreement would be:

$$P_e = \frac{1}{k} \qquad \textbf{Eqn. 1}$$

where in our case $k$ would be 4.

An alternative definition of chance agreement considers that the groups' proclivity to distribute their ratings in a certain way is a source of disagreement:

$$P_e = \sum_{i=1}^{4} P_{i+} P_{+i} \qquad \textbf{Eqn. 2}$$

The above marginal proportions are maximum likelihood estimates of the population proportions under a multinomial sampling model. If each of the assessors makes ratings at random according to the marginal proportions, then the above is chance agreement (derived using the multiplication rule of probability and assuming independence between the two groups).

A general form for agreement coefficients is [172]:

$$Agreement = \frac{P_O - P_e}{1 - P_e}$$

The observed agreement that is in excess of chance agreement is given by $P_O$ - $P_e$. The maximum possible excess over chance agreement is 1 - $P_e$. Therefore, this type of agreement coefficient is the ratio of observed excess over chance agreement to the maximum possible excess over chance agreement.

When there is complete agreement between the two groups, $P_O$ will take on the value of 1. In this case, the agreement coefficient is 1. If observed agreement is greater than chance, then the agreement coefficient is greater than zero. If observed agreement is less than would be expected by chance, then the agreement coefficient is less than zero.

An agreement coefficient[32] that considers chance agreement as in Eqn. 1 is Bennett et al.'s S coefficient [10]. An agreement coefficient that considers chance agreement as in Eqn. 2 is Cohen's Kappa ($\kappa$) [28].

A priori it would seem more reasonable to assume that each group has a proclivity to distribute their ratings in a certain way rather than assume that both groups distribute their ratings in exactly the same way. This therefore suggests that Cohen's Kappa is a more appropriate coefficient. Furthermore, as Cohen notes, it is also desirable to take account of disagreement in marginal distributions in an agreement coefficient [28].[33]

Cohen's coefficient Kappa ( $\kappa$) is therefore defined as:[34]

---

[32] It should be noted that "agreement" is different from "association". For the ratings from two teams to agree, the ratings must fall in the same adequacy category. For the ratings from two teams to be associated, it is only necessary to be able to predict the adequacy category of one team from the adequacy category of the other team. Thus, strong agreement requires strong association, but strong association can exist without strong agreement. For instance, the ratings can be strongly associated and also show strong disagreement.

[33] A further problem with the S coefficient is that it is dependent on the number of categories in a rating scheme [150]. For example, the chance agreement for a two category rating scheme is 0.5, while chance agreement for a four category rating scheme is 0.25. Therefore, by definition, the S coefficient rewards rating schemes with larger numbers of categories. While this is not a major problem when the number of categories is fixed, as is the case with ISO/IEC 15504 based assessments, it provides another reason for not using this coefficient in general.

[34] There is also a weighted version of the Kappa coefficient [29]. The weighting schemes that have been suggested are based on mathematical convenience. However, thus far no weighting scheme has been developed that has a substantive meaning in the context of software process assessments.

$$\kappa = \frac{P_O - P_e}{1 - P_e}$$

where the definition $P_e$ is as in Eqn. 2.

The minimum value of $\kappa$ depends upon the marginal proportions. However, since we are interested in evaluating agreement, the lower limit of $\kappa$ is not of interest.

### 7.3.1 Example Kappa Calculation

Here we illustrate the computation of the Kappa coefficient through an example. Consider Table 11, which contains the proportions calculated from a hypothetical assessment.

| | | **Group 1** | | | | |
|---|---|---|---|---|---|---|
| | | **F** | **L** | **P** | **N** | |
| **Group 2** | **F** | 0.051 | 0.128 | 0.077 | 0 | 0.256 |
| | **L** | 0 | 0.128 | 0.230 | 0.026 | 0.384 |
| | **P** | 0 | 0.026 | 0.154 | 0.128 | 0.307 |
| | **N** | 0 | 0 | 0 | 0.051 | 0.051 |
| | | 0.051 | 0.282 | 0.461 | 0.205 | |

**Table 11:** 4x4 table containing proportions from a hypothetical assessment.

For this particular table we have:

$$P_O = 0.051 + 0.128 + 0.154 + 0.051 = 0.384$$

$$P_e = (0.256 \times 0.051) + (0.384 \times 0.282) + (0.307 \times 0.461) + (0.051 \times 0.205) = 0.273$$

$$\kappa = \frac{0.384 - 0.273}{1 - 0.273} = 0.153$$

### 7.3.2 Interrater Agreement Results

A recent study that summarized the results from the interrater agreement studies that have been performed produced the benchmark shown in Table 12 [68]. This indicates the quartile values for the Kappa coefficient from actual studies. Therefore, 25% of the assessed ISO/IEC 15504 process instances has Kappa values below or equal to 0.44, and 25% had values greater than 0.78.

| Kappa Statistic Range | Strength of Agreement | Percentile Interpretation | |
|---|---|---|---|
| $\kappa \leq 0.44$ | Poor | (bottom 25%) | (bottom 50%) |
| $0.44 < \kappa \leq 0.62$ | Moderate | | |
| $0.62 < \kappa \leq 0.78$ | Substantial | | (top 50%) |
| $\kappa > 0.78$ | Excellent | (top 25%) | |

**Table 12:** ISO/IEC 15504 Software Process Assessment Kappa benchmark.

It is clear that variation in interrater agreement exists, and the factors that may be causing such variation are discussed below. In order to determine whether these values are indicative in general of a reasonable amount of reliability or not, it is informative to compare this benchmark with one used in medicine to evaluate the reliability of doctors'diagnoses.

| Kappa Statistic | Strength of Agreement |
|---|---|
| <0.40 | Poor |
| 0.40-0.75 | Intermediate to Good |
| >0.75 | Excellent |

**Table 13:** The Fleiss Kappa benchmark.

The benchmark provided by Fleiss [76] is shown in Table 13. We can see that what is characterized as poor diagnosis reliability falls in the lowest quartile on the assessment benchmark. Therefore, at least by these standards, the majority of process assessments have a respectable level of interrater agreement.

## 7.4 Factors Affecting Reliability

A survey was conducted to prioritize the factors that have an impact on the reliability of process assessments [59]. The data collection was conducted during a meeting of the SPICE project that took place in Mexico in October 1996. These meetings are of sizeable number of experienced assessors with substantial experience in various assessment methods and models, such as the CMM, CBA IPI, TRILLIUM, BOOTSTRAP, and other proprietary models and methods. During the meeting, the authors generated a list of factors that may potentially have an impact on the reliability of assessments. The authors relied largely on their experiences and the prior comments of other assessors. This is justifiable given that no comprehensive study of the factors influencing reliability had been conducted.

This list was reviewed by two other experienced assessors to ensure completeness of coverage. The list is given in Figure 17. The refined list was turned into a rating form. The rating form was piloted with 4 assessors to ensure that it was understandable and to identify ambiguities. Based on this feedback, a new form was developed and was distributed to all attendees at the closing session of the meeting. In total, approximately 50 individuals attended the project meeting, and we expect a slightly smaller number attended the closing session. A total of 26 valid responses were received back and were used for analysis.

The form consisted of an unordered list of factors that are believed to have an impact on the reliability of assessments. The respondent was requested to rate each factor on a five point scale, where 1 means that the factor has "very high influence" on the reliability of assessments, and 5 means that it has "very low influence". The objective was to prioritize these factors. The responses on the 5-point scale were dichotomized into HIGH INFLUENCE (scores 1 and 2) and LOW INFLUENCE (scores 3 to 5). For each factor, the percentage of respondents who rated a factor as HIGH INFLUENCE was calculated. This percentage is used for ranking.

The results are presented in Figure 17. We use the letters A to W in the discussions to indicate items in Figure 17.

| Id | Factor | |
|---|---|---|
| A | Lead assessor's experience/competence in conducting assessments | (24/26) = 92% |
| B | Lead assessor's knowledge of ISO/IEC 15504 documents | (22/25) = 88% |
| C | Clarity of the semantics of the process definition in the ISO/IEC 15504 documentation | (22/26) = 84.6% |
| D | The extent to which the assessment process is defined and documented | (20/26) = 77% |
| E | Team members' knowledge of ISO/IEC 15504 documents | (16/25) = 64% |
| F | Amount of collected data (objective evidence and/or interviews) | (16/26) = 61.5% |
| G | Assessees' commitment | (13/25) = 52% |
| H | Assessment team stability | (13/25) = 52% |
| I | Rating just after collecting the evidence, and validation at the end of the assessment | (12/25) = 48% |
| J | Assessment team composition (unidisciplinary vs. multidisciplinary competencies) | (11/25) = 44% |
| K | Sponsor commitment | (11/25) = 44% |
| L | Team building curve | (10/25) = 40% |
| M | Competence of the interviewed assessees | (10/25) = 40% |
| N | Number of assessed projects in the organizational unit | (9/25) = 36% |
| O | Assessment duration | (9/25) = 36% |
| P | Lead assessor's experience/competence in conducting audits | (8/25) = 32% |
| Q | Assessment team size (number of assessors including lead assessor) | (8/25) = 32% |
| R | Rating only at the end of the assessment | (8/25) = 32% |
| S | Language used during the assessment | (8/25) = 32% |
| T | Time allocation between artifact reviews and interviews | (8/25) = 32% |
| U | Management of the assessment logistics (e.g., availability of facilities) | (6/25) = 24% |
| V | The capability of the organizational unit's processes | (5/25) = 20% |
| W | Whether the assessors are external or internal | (3/26) = 11.5% |

**Figure 17:** Ranking of factors affecting the reliability of assessments.

### 7.4.1 Assessor Competence

Given that assessment are a subjective measurement procedure, it is expected that assessor competence will have an impact on the reliability of assessments. We consider mainly the competence of the lead assessor since s/he is the key person on the assessment team. The types of competencies covered here include knowledge of the ISO/IEC 15504 documents (B), experience and competence in conducting assessments (A) and audits (P). Indeed, a subsequent study demonstrated that assessors lacking experience can give results that are quite different from experienced assessors [115]. Also, the knowledge of the ISO/IEC 15504 documents (E) was perceived to be important since the team members will be collecting, organizing, and interpreting information during an assessment, they must know ISO/IEC 15504 well to collect the right information, organize it efficiently, and interpret it properly.

### 7.4.2 External vs. Internal Assessors

Previous research has identified potential systematic biases of internal or external assessors [57] (i.e., one assessor would systematically rate higher or lower than the other). For example, an internal assessor may favor the organization in his/her ratings or may have other information not available to the external assessor which may influence the ratings. Similarly, an external assessor may not know the organization's business well and therefore may systematically underrate the implementation of its practices. This issue is covered in item (W).

### 7.4.3 Team Size

Practice and recommendation on team size have tended to be confusing. In some assessment methods it is stipulated that teams range in size from 5 to 9 [49]. In the first version of the ISO/IEC 15504 documents the recommendation has been team sizes of at least two assessors. From a practical point of view it has been suggested that a single assessor would find it difficult to collect and record information at the same time, and therefore more than one person is recommended. Item (Q) covers this issue from the perspective of its impact on reliability.

### 7.4.4 Backgrounds of Assessors

It has been noted by some assessors that multidisciplinary assessment teams (i.e., not consisting of only software engineering staff, but also those with backgrounds in, for example, human resources management and marketing) are better able to collect the right evidence (i.e., ask the right questions and request the appropriate documents) and better able to interpret it for certain processes. This would likely increase the reliability of assessments. This issue is covered in item (J).

### 7.4.5 Number of Assessed Projects

During an assessment, a sample of projects is selected for assessment. It is usually not feasible to assess all of the projects within the scope of the organization. It is assumed, through this selection process, that the selected projects are representative of the whole organization. Clearly, the more projects that are assessed, the more representative and hence repeatable the ratings that are made. This is covered in item (N). Of course, this item applies only when one is giving ratings to whole organizations, and has less influence when the unit of analysis is a process instance.

### 7.4.6 Assessment Duration

Long assessment may lead to fatigue of the assessors and assessees, may reduce their motivation, and hence reduce the reliability of ratings. Short assessments may not collect sufficient information to make reliable ratings. This is covered in item (O).

### 7.4.7 Team Building Curve

In team-based assessments, it is expected that the assessor judgements would converge as the assessment progresses. This would be due to a better appreciation of the other team members' experiences, backgrounds, and due to the consensus building activities that usually take place during an assessment. This is covered in L.

### 7.4.8 Clarity of Documents

Ambiguities and inconsistencies in the definition of practices or in the scales used to make ratings would potentially lead to different interpretations of what practices mean and how to rate them. This would in turn reduce reliability. This issue is covered in C.

### 7.4.9 Definition of the Assessment Process

Having a clearly defined assessment process potentially ensures that the process is repeatable, which in turn has an impact on the repeatability of ratings. This is covered in D.

### 7.4.10 Amount of Data Collected

The more time spent on data collection (F), the more data will be collected. The more data that is collected, the more likely that the assessment team will have a more objective basis to make their ratings.

Furthermore, the extent to which time is allocated to different methods for data collection may have an impact on the amount of data collected (T).

#### 7.4.11 Capability of Organization and its Processes

It is hypothesized that higher capability processes are easier to rate because of the existence of more objective evidence and process stability to make consistent judgements. This is covered in U.

#### 7.4.12 Assessment Method

A feature of the assessment method is when the ratings are actually made. One approach is to collect data about a process and then make the ratings right afterwards (I). Another approach is to collect data on all of the processes within the scope of the assessment, and then rate them all afterwards (R). The latter allows the assessors to build an overall picture of the implementation of software engineering practices in the organization, and also to get a better understanding of the organization's business and objectives (especially for external assessors) before making ratings. This could potentially increase the reliability of assessments.

One study investigated the impact of the method on the agreement among independent assessors in their ratings [59]. The results indicate that for low capability levels, there is a difference in reliability between rating processes early in assessment versus late in an assessment. For higher capability processes, it does not make a difference whether ratings are done early or late in an assessment.

#### 7.4.13 Sponsor and Assessee Commitment

A lack of commitment by members of the assessed organization can lead to insufficient or inappropriate resources being made available for the assessment. This may compromise the assessment team's ability to make repeatable ratings. This issue is covered in items (K and G).

#### 7.4.14 Assessment Team Stability

If the assessment team changes during an assessment, the disruption can break the consensus building cycle. Furthermore, knowledge about the organization that has been gained by an assessor that leaves would have to be regained by a new assessor. This is covered in item H.

#### 7.4.15 Logistics Management

Inappropriate management of the logistics may distract the assessors and waste time. This could potentially lead to insufficient evidence being collected and hence to lower reliability. This issue is covered in (V).

#### 7.4.16 Assessee Competence

Assessees provide the necessary information during an assessment. If the assessees are not competent then they may provide inconsistent information to the assessors, which may consequently lead to inconsistent interpretations of the process'capability. This issue is covered in item (M).

#### 7.4.17 Assessment Language

Assessments are now being conducted all around the world. In fact, in the first phase of the SPICE trials certain documents were translated to a language other than English. The issue of the impact of language on the reliability of assessments is covered in (S).

#### 7.4.18 Discussion

The results clearly indicate that assessment team competence and the clarity of the documents are perceived to be the two most important factors that have an impact on the reliability of assessments.

Equally interesting are the factors that were rated to be of least priority. This does not mean that they are not important, only that they are less important than the other factors. These factors were whether the assessors were internal vs. external, the capability of assessed processes, and the assessment logistics.

## 7.5 Summary

In summary, the following is what we have learned thus far about the reliability of process assessments:

- The reliability of SPAs is important to give confidence in the decisions based on assessment results, but also a recent study also found evidence that more reliable assessments are less costly. The reason for that is when assessments are reliable (i.e., the assessors agree in their initial ratings), the consolidation phase progresses faster with less consensus building, hence resulting in an overall reduction in cost.

- A number of studies have evaluated the internal consistency of assessment instruments. However, most of these assumed unidimensionality, which, as noted above, was found subsequently not to be the case. One study evaluated the internal consistency of the ISO/IEC 15504 capability scale dimensions and found it to be sufficiently high for practical usage.

- Studies of interrater agreement in ISO/IEC 15504 based assessments indicate that most ratings by independent teams are sufficiently reliable (in at least 75% of studied ratings).

- Current evidence demonstrates that the reliability of SPAs deteriorates with inexperienced assessors (the sole study compared inexperienced with experienced assessor ratings during a training course).

- In reliability studies with both internal and external assessors on an assessment team, systematic bias was witnessed in some cases. Assessment methods should be designed to alleviate this possibility.

- It has been suggested by assessment experts in a survey that the following are the most important factors that affect the reliability of assessments:
  - Clarity of the best practice model, and knowledge of it by the assessment team
  - The extent to which the assessment process is defined and documented
  - Amount of data that is collected during the assessment
  - Assessee and sponsor commitment to the assessment
  - Assessment team composition and stability

However, many of these factors require further systematic empirical investigation.

- It was found that the assessment method has an impact on the reliability of assessments (interrater agreement). A feature of the assessment method is when the ratings are actually made. One type of method stipulates that the assessment team collect data about a process and then make the ratings right afterwards. Another method would be to collect data on all of the processes within the scope of the assessment, and then rate them all afterwards. The latter allows the assessors to build an overall picture of the implementation of software engineering practices in the organization, and also to get a better understanding of the organization's business and objectives (especially for external assessors) before making ratings. One study found that for low capability levels, there is a difference in reliability between rating processes early in assessment versus late in an assessment. For higher capability processes, it does not make a difference whether ratings are done early or late in an assessment, but for low capability processes late ratings are more reliable.

# 8 The Validity of Process Capability Measures

## 8.1 Types of Validity

The validity of measurement is defined as the extent to which a measurement procedure is measuring what it is purporting to measure [114]. During the process of validating a measurement procedure one attempts to collect evidence to support the types of inferences that are to be drawn from measurement scores [33]. In the context of SPAs, concern with validity is epitomized by the question "are assessment ratings really measuring best software process practices ?".

A basic premise of SPAs is that the resultant quantitative ratings are associated with the performance of the project and/or organization that is assessed. This premise consists of two parts:[35]

- that the practices defined in the best practice model are indeed good practices and their implementation will therefore result in improved performance
- that the quantitative assessment rating is a true reflection of the extent to which these practices are implemented in the organization or project; and therefore projects or organizations with higher assessment ratings are likely to perform better.

Validity is related to reliability in the sense that reliability is a necessary but insufficient condition for validity. The differences between reliability and validity are illustrated below by way of two examples.

For instance, assume one seeks to measure intelligence by having children throw stones as far as they could. The distance the stones are thrown on one occasion might correlate highly with how far they are thrown on another occasion. Thus, being repeatable, the stone-throwing measurement procedure would be highly reliable. However, the distance that stones are thrown would not be considered by most informed observers to be a valid measure of intelligence.

As another example, consider a car's fuel gauge that systematically shows five liters higher than the actual level of fuel in the gas tank. If repeated readings of fuel level are taken under the same conditions, the gauge will yield consistent (and hence reliable) measurements. However, the gauge does not give a valid measure of fuel level in the gas tank.

We consider here two types of validity that we believe are most important for SPAs: content and predictive validity.

### 8.1.1 Content Validity

Content validity is defined as the representativeness or sampling adequacy of the content of a measuring instrument [114]. Ensuring content validity depends largely on expert judgement.

In the context of SPAs, expert judgement would ensure that assessments are at least perceived to measure best software engineering practice. This centers largely on the content of the models.

At least for the best practice models that we are intimately associated with, the models have been extensively reviewed by experts in industry and academe, and their feedback has been accounted for in the revision of these models. This exercise, coupled with the feedback obtained from actual field applications of the models, ensures to a certain degree that the models adequately cover the content domain.

To further ensure content validity, it is necessary that all assessment instruments include questions that adequately sample from the content domain. However, this requirement is easily met since most assessment instruments are derived directly from the best practice models.

### 8.1.2 Predictive Validity

A predictive validity study typically tests for a relationship between process capability and performance. This relationship is expected to be dependent upon some context factors (i.e., the relationship functional form or direction may be different for different contexts, or may exist only for some contexts).

The hypothesized model can be tested for different units of analysis [83]. The three units of analysis are the life cycle process (e.g., the design process), the project (which could be a composite of the capability of multiple life cycle processes of a single project, such as design and coding), or the organization (which could be a composite of the capability of the same or multiple processes across different projects). All of the three variables in the model can be measured at any one of these units of analysis.

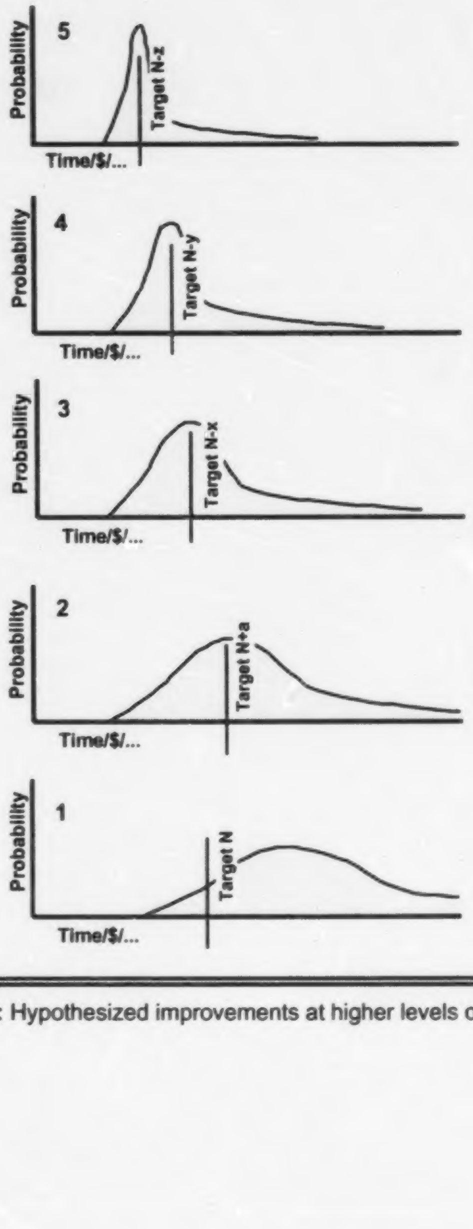
---

[35] The fact that the premise behind the use of quantitative scores from SPAs consists of two parts means that if no empirical evidence is found to support the basic premise, then we would not know which part is at fault. For example, if we find that there is no relationship between the assessment score and performance it may be because:

- the practices are really not good practices, but the measurement procedure is accurately measuring their implementation, or
- the practices are really good practices, but the measurement procedure is not accurately measuring their implementation.

From a practical standpoint it does not matter which of the above two conclusions one draws since the practices and measurement procedure are always packaged and used together.

The literature refers to measures at different units of analysis using different terminology. To remain consistent, we will use the term "process capability" and preface it with the unit of analysis where applicable. For example, one can make a distinction between measuring process capability, as in ISO/IEC 15504, and measuring organizational maturity, as in the SW-CMM [137]. Organizational maturity can be considered as a measure of organizational process capability.



Performance continually improves in Level 5 organizations.

Based on quantitative understanding of process and product, performance continues to improve in Level 4 organizations.

With well-defined processes, performance improves in Level 3 organizations.

Plans based on past performance are more realistic in Level 2 organizations.

Schedule and cost targets are typically overrun by Level 1 organizations.

**Figure 18:** Hypothesized improvements at higher levels of SW-CMM capability (source [139]).

## 8.2 Predictive Validity Hypotheses[36]

The basic hypotheses related to predictive validity can be explained with reference to Figure 18 (see [139]). The first improvement that is expected when process capability increases is in predictability. The difference between targeted results and actual results is expected to decrease. The second expected improvement is in control. The variation in actual results around target results gets narrower. The third expected improvement is in effectiveness. The actual targeted results improve as process capability increases. The hypotheses can be stated with respect to project performance or organizational performance.

The predictive validity studies that have been performed focused on the first and third issues. Demonstrating a reduction in variation has, to our knowledge, not been studied.

## 8.3 Validation Approaches

Two classes of empirical studies have been conducted and reported thus far: case studies and correlational studies [83]. Case studies describe the experiences of a single organization (or a small number of selected organizations) and the benefits it gained from increasing its process capability. Case studies are most useful for showing that there are organizations that have benefited from increased process capability. Examples of these are reported in [102][92][40][41][171][11][124][22][122] (also see [118] for a recent review). However, in this context, case studies have a methodological disadvantage that makes it difficult to generalize the results from a single case study or even a small number of case studies. Case studies tend to suffer from a selection bias because:

- Organizations that have not shown any process improvement or have even regressed will be highly unlikely to publicize their results, so case studies tend to show mainly success stories (e.g., all the references to case studies above are success stories), and
- The majority of organizations do not collect objective process and product data (e.g., on defect levels, or even keep accurate effort records). Only organizations that have made improvements and reached a reasonable level of maturity will have the actual objective data to demonstrate improvements (in productivity, quality, or return on investment). Therefore failures and non-movers are less likely to be considered as viable case studies due to the lack of data.

With correlational studies, one collects data from a larger number of organizations or projects and investigates relationships between process capability and performance statistically. Correlational studies are useful for showing whether a general association exists between increased capability and performance, and under what conditions.

| | | **Measuring the Criterion** | | |
|---|---|---|---|---|
| | | Questionnaire | Measurement Program | |
| **Measuring Capability** | **Questionnaire** | Q1 | Q2 | (low cost) |
| | **Assessment** | Q3 | Q4 | (high cost) |
| | | (across organizations) | (within one organization) | |

**Table 14:** Different correlational approaches for evaluating predictive validity.

Correlational approaches to evaluating the predictive validity of a process capability measure can be classified by the manner in which the variables are measured. Table 14 shows a classification of approaches. The columns indicate the manner in which the criterion is measured. The rows indicate the manner in which the process capability is measured. The criterion can be measured using a questionnaire whereby data on the perceptions of experts are collected. It can also be measured through a measurement program. For example, if our criterion is defect density of delivered software products, then this could be measured through an established measurement program that collects data from defects found in the field. Process capability can also be measured through a questionnaire whereby data on the perceptions of experts on the capability of their processes are collected. Alternatively, actual assessments can be performed, which are a more rigorous form of measurement[37].

---

[36] Material in this section is based partially on [66][67].

A difficulty with studies that attempt to use criterion data that are collected through a measurement program is that the majority of organizations do not collect objective process and product data (e.g., on defect levels, or even keep accurate effort records). Primarily organizations that have made improvements and reached a reasonable level of process capability will have the actual objective data to demonstrate improvements (in productivity, quality, or return on investment). This assertion is supported by the results in [16] where, in general, it was found that organizations at lower SW-CMM maturity levels are less likely to collect quality data (such as the number of development defects). Also, the same authors found that organizations tend to collect more data as their CMM maturity levels rise. It was also reported in another survey [147] that for 300 measurement programs started since 1980, less than 75 were considered successful in 1990, indicating a high mortality rate for measurement programs. This high mortality rate indicates that it may be difficult right now to find many organizations that have implemented measurement programs.

This means that organizations or projects with low process capability would have to be excluded from a correlational study. Such an exclusion would reduce the variation in the performance measure, and thus reduce (artificially) the validity coefficients. Therefore, correlational studies that utilize objective performance measures are inherently in greater danger of not finding significant results.

Furthermore, when criterion data are collected through a measurement program, it is necessary to have the criterion measured in the same way across all observations. This usually dictates that the study is done within a single organization where such measurement consistency can be enforced, hence reducing the generalizability of the results.

Conducting a study where capability is measured through an assessment as opposed to a questionnaire implies greater costs. This usually translates into smaller sample sizes and hence reduced statistical power. Therefore, the selection of a quadrant in Table 14 is a tradeoff among cost, measurement rigor, and generalizability.

Many previous studies that evaluated the relationship between process capability (or organizational maturity) and the performance of projects tended to be in quadrant Q1. For example, [83][39][25]. These studies have the advantage that they can be conducted across multiple projects and across multiple organizations, and hence can produce more generalizable conclusions.

A more recent study evaluated the relationship between questionnaire responses on implementation of the SW-CMM KPA's and defect density [119], and this would be placed in quadrant Q2. However, this study was conducted across multiple projects within a single organization, reducing its generalizability compared with studies conducted across multiple organizations.

The ISO/IEC 15504 studies in [66][67] can be placed in quadrant Q3 since the authors use process capability measures from actual assessments, and questionnaires for evaluating project performance. This retains the advantage of studies in quadrant Q1 since it is conducted across multiple projects in multiple organizations, but utilizes a more rigorous measure of process capability. Similarly, the study of Jones can be considered to be in this quadrant [107][108]. [38]

---

[37] More rigorous"is intended to mean with greater reliability and construct validity.

[38] Since it is difficult to find low maturity organisations with objective data on effort and defect levels, and since there are few high maturity organisations, Jones' data relies on the reconstruction of, at least, effort data from memory, as noted in [109]: "The SPR approach is to ask the project team to reconstruct the missing elements from memory." The rationale for that is stated as "the alternative is to have null data for many important topics, and that would be far worse." The general approach is to show staff a set of standard activities, and then ask them questions such as which ones they used and whether they put in any unpaid overtime during the performance of these activities. For defect levels, the general approach is to do a matching between companies that do not measure their defects with similar companies that do measure, and then extrapolate for those that don't measure. It should be noted that SPR does have a large data base of project and organisational data, which makes this kind of matching defensible.

Studies in quadrant Q4 are likely to have the same limitations as studies in quadrant Q2: being conducted across multiple projects within the same organization. For instance, the study of McGarry et al was conducted within a single company [130], and the AFIT study was conducted with contractors of the Air Force [77][121].

Therefore, the different types of studies that can be conducted in practice have different advantages and disadvantages, and predictive validity studies have been conducted in the past that populate all four quadrants. It is reasonable then to encourage studies in all four quadrants. Consistency in the results across correlational studies that use the four approaches would increase the weight of evidence supporting the predictive validity hypothesis.

## 8.4 Main Effects

There have been a few correlational studies in the past that evaluated the predictive validity of various process capability measures. For example, Goldenson and Herbsleb [83] evaluated the relationship between SW-CMM capability ratings and organizational performance measures. They surveyed individuals whose organizations have been assessed against the SW-CMM. The authors evaluated the benefits of higher process capability using subjective measures of performance. Organizations with higher capability tend to perform better on the following dimensions (respondents chose either the "excellent" or "good" response categories when asked to characterize their organizations performance on these dimensions): ability to meet schedule, product quality, staff productivity, customer satisfaction, and staff morale. The relationship with the ability to meet budget commitments was not found to be statistically significant.

A more recent study considered the relationship between the implementation of the SW-CMM KPAs and delivered defects (after correcting for size and personnel capability) [119]. They found evidence that increasing process capability is associated with fewer delivered defects. Another correlational study investigated the benefits of moving up the maturity levels of the SW-CMM [77][121] (also see the reanalysis of the data from this study in the appendix, Section 10). They obtained data from historic U.S. Air Force contracts. Two measures were considered: (a) cost performance index which evaluates deviations in actual vs. planned project cost, and (b) schedule performance index which evaluates the extent to which schedule has been over/under-run. Generally, the results show that higher maturity projects approach on-target cost, and on-target schedule. McGarry et al. [130] investigated the relationship between assessment ratings using an adaptation of the SW-CMM process capability measures and project performance for fifteen projects within a single organization. They did not find strong evidence of predictive validity, although they were all in the expected direction. Clark [25] investigated the relationship between satisfaction of SW-CMM goals and software project effort, after correcting for other factors such as size and personnel experience. His results indicate that the more KPAs are implemented, the less effort is consumed on projects. Jones presents the results of an analysis on the benefits of moving up the 7-level maturity scale of Software Productivity Research (SPR) Inc.'s proprietary model [107][108]. This data were collected from SPR's clients. His results indicate that as organizations move from Level 0 to Level 6 on the model they witness (compound totals): 350% increase in productivity, 90% reduction in defects, 70% reduction in schedules.

Deephouse et al. evaluated the relationship between individual processes and project performance [39]. As would be expected, they found that evidence of predictive validity depends on the particular performance measure that is considered. One study by El Emam and Madhavji [58] evaluated the relationship between four dimensions of organizational process capability and the success of the requirements engineering process. Evidence of predictive validity was found for only one dimension.

---

However, since at least some of the criterion measures are not collected from measurement programs, we place this study in the same category as those that utilise questionnaires.

| Performance Measure | Process(es) |
|---|---|
| **Small Organizations** | |
| Ability to meet budget commitments | |
| Ability to meet schedule commitments | Develop Software Design |
| Ability to achieve customer satisfaction | |
| Ability to satisfy specified requirements | |
| Staff productivity | |
| Staff morale / job satisfaction | |
| **Large Organizations** | |
| Ability to meet budget commitments | Develop Software Design<br>Implement Software Design |
| Ability to meet schedule commitments | Develop Software Design |
| Ability to achieve customer satisfaction | Develop Software Design |
| Ability to satisfy specified requirements | Develop Software Design |
| Staff productivity | Develop Software Requirements<br>Integrate and Test Software |
| Staff morale / job satisfaction | Develop Software Design |

**Table 15:** Summary of the findings from the predictive validity study. In the first column are the performance measures that were collected for each project. In the second column are the development processes whose capability was evaluated. The results are presented separately for small (equal to or less than 50 IT staff) and large organizations (more than 50 IT Staff).

The results from recent studies that evaluate the predictive validity of the process capability measures in ISO/IEC 15504 [66][67] are shown in Table 15. In general, these indicate that process capability for some processes is related to performance, but mainly only for large organizations. The evidence for small organizations is rather weak.

Many software organizations are being assessed against the clauses of ISO 9001. A number of surveys have been conducted that evaluate the benefits of ISO 9001 registration in industry in general and in software organizations in particular. Some of the results of these surveys have been presented in [164]. Below we summarize some of the relevant findings:

- One survey conducted in 1993 had 292 responses with almost 80% of the responding organizations being registered to ISO 9001. The findings included:
  - 74% felt that the benefits of registration outweighed the costs
  - 54% received favorable feedback from their customers after registration
- A survey of companies in the U.K. had 340 responses from companies that were registered. It was found that 75% of the respondents felt that registration to ISO 9001 improved their product and/or service.
- A survey of companies that were registered in the U.S.A. and Canada with 620 responses found that:

  - the most important internal benefits to the organization included: better documentation (32.4%), greater quality awareness (25.6%), a positive cultural change (15%), and increased operational efficiency/productivity (9%); and
  - the most important external benefits to the organization included: higher perceived quality (33.5%), improved customer satisfaction (26.6%), gaining a competitive edge (21.5%), and reduced customer quality audits (8.5%).
- A survey of 45 software organizations in Europe and North America that have become ISO 9001 registered found that:
  - 26% reported maximum benefit from increased efficiency
  - 23% reported maximum benefit from increased product reliability
  - 22% reported maximum benefit from improved marketing activity
  - 14% reported maximum benefit from cost savings, and
  - 6% reported maximum benefit from increases exports

Thus, with respect to registration to ISO 9001, the few studies that have been conducted are consistent in their findings of benefits to registration. However, many of these studies were not specific to software organizations. Therefore, more research specifically with software organizations would help the community better understand the effects of registration.

## 8.5 Moderating Effects

A recent article noted that existing evidence suggests that the extent to which a project's or organization's performance improves due to the implementation of good software engineering practices (i.e., increasing process capability) is dependent on the context [63]. This highlights the need to consider the project and/or organizational context in predictive validity studies. However, it has also been noted that the overall evidence remains equivocal as to which context factors should be considered in predictive validity studies [63].

One of the important moderating variables that has been mentioned repeatedly in the literature is organizational size.

Previous studies provide inconsistent results about the effect of organizational size. For example, there have been some concerns that the implementation of some of the practices in the CMM, such as a separate Quality Assurance function and formal documentation of policies and procedures, would be too costly for small organizations [15]. Therefore, the implementation of certain processes or process management practices may not be as cost-effective for small organizations as for large ones. However, a moderated analysis of the relationship between organizational capability and requirements engineering process success (using the data set originally used in [53]) [63] found that organizational size does not affect predictive validity. This result is consistent with that found in [83] for organization size and [39] for project size, but is at odds with the findings from [15][66][67].

To further confuse the issue, an earlier investigation [123] studied the relationship between the extent to which software development processes are standardized and MIS success.[39] It was found that standardization of life cycle processes was associated with MIS success in smaller organizations but not in large ones. This is in contrast to the findings cited above. Therefore, it is not clear if and how organization size moderates the benefits of process and the implementation of process management practices.

## 8.6 Diminishing Rates of Return

Some studies suggest that there may be diminishing returns with greater process capability. For example, the functional form of the relationship between the SW-CMM based maturity measure and field defects as reported in [119] is shown in Figure 19. As can be seen there, the number of defects decreases as process capability increases. However, the rate of the decrease diminishes and further improvements in process capability produce only marginal improvements in field defects.[40]

---

[39] Process standardisation is a recurring theme in process capability measures.

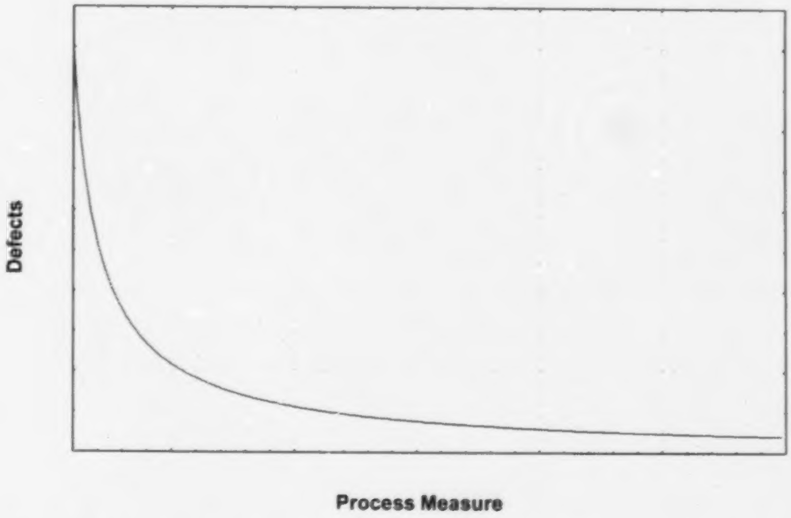


**Figure 19:** Functional form of the relationship between the process measure and number of defects found (from [119]). We do not show the values on the axes since they would be meaningless without describing in detail how process capability was measured. This is not necessary here, since we are only interested in showing the functional form of the relationship.

A similar interpretation can be made based on the results of Clark's analysis [25]. In that study, the process measure was coded so that smaller values mean higher capability. Larger values mean smaller capability. The functional form of the relationship in that study is shown in Figure 20. As process capability increases, person-months spent on the projects decrease. That is, there tends to be greater productivity. Similar to the previous study, however, the rate of decrease goes down as process capability increases.[41]

---

[40] The authors controlled for the size of the product in their analysis, so the graph can be interpreted as applicable regardless of product size.

[41] Also similar to the other study, the model adjusted statistically for the effect of size and other covariates

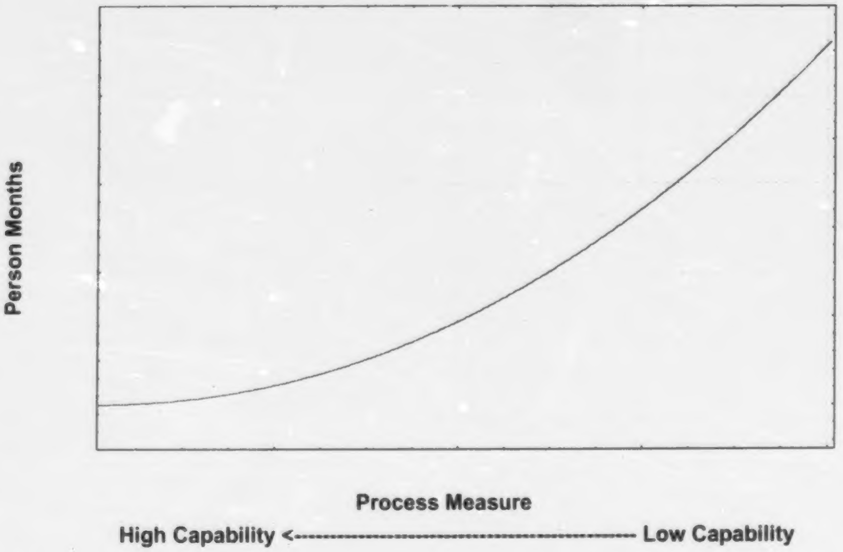


**Figure 20:** Functional form of the relationship between the process measure and person months (from [25]). We do not show the values on the axes since they would be meaningless without describing in detail how process capability was measured. This is not necessary here, since we are only interested in showing the functional form of the relationship.

These results do suggest that there may be diminishing returns at higher levels of process capability, at least with respect to product defects and development effort. However, the results should not be overinterpreted. There are no software projects that take zero time to complete, and the proportion of software products in the world which originally shipped with zero defects is likely rather small. Such ceiling/floor effects, associated with the measures of effort and quality investigates, may appear in the form of asymptotic limits in the relationships studied.

Depending on an organization's quality goals, there is still room for improvement until one reaches zero defects. Continuing attention to process capability may be necessary to maintain such pattterns over time, particularly if the remaining defects may differ in severity. Moreover, at higher levels of capability, organizations may choose to focus on other aspects of process and product improvement. The point where the rate of return is "not worth the additional effort" remains an empirical question .

## 8.7 Causality

None of the studies reviewed establishes a causal relationship, i.e., that the changes in performance are caused by the change in process capability. To establish causation one must at least rule out other possible causal factors that could have led to the performance changes (also, experience reports documenting benefits of SPI would have to rule out natural progress, i.e., if the organization did not make any changes, would it have achieved the same benefits?).

It is clear that implementation of processes or process management practices are not the only factors that will influence performance. Bach [1] has made the argument that individual software engineer capabilities is a critical factor having an impact on project and organizational effectiveness. He even goes further, stating *"that the only basis for success of any kind is the 'heroic efforts of a dedicated team'."* The importance of individual capability is supported by empirical research. For instance, one study found that the capabilities of the lead architect were related to the quality of requirements engineering products [51]. Another study found a relationship between the capability of users participating in the requirements engineering process and its success [70]. Other field studies of requirements and design processes also emphasized the importance of individual capabilities [34][54].

The implementation of automated tools has been advocated as a factor that has an impact on effectiveness. This assertion is supported by empirical research. For instance, one study of the implementation of an Information Engineering toolset achieved increases in productivity and decreases in post-release failures [75].

The best that can be attained with studies that focus only on process factors is strong evidence that process capability is *associated* with performance or that organizations *could* benefit from SPI activities. In order to improve our understanding of the influences of other factors on performance more sophisticated empirical studies would have to be conducted. These would include building multivariate models that take the influence of non-process factors into account and investigate the interactions between process and non-process factors. Good current examples of these are the studies in[25][119].

## 8.8 Summary

In summary, we can make the following general statements:

- There is ample evidence through case studies that higher process capability is associated with improved performance. These demonstrate that it is plausible to improve performance as capability is improved.
- The results of more methodologically defensible predictive validity studies of capability measures tend to demonstrate an association between increased process capability and increased performance. Performance was measured both at the project and organizational levels. No evidence exists to our knowledge that demonstrates a reduction in variability in performance as process capability increases.
- Current studies indicate potential diminishing rates of return to increased process capability. It is not yet clear whether this means that there exists a maximal gain from SPI based on contemporary best practice models or whether the effect observed in the studies is a statistical artifact. This is a topic that should be further investigated in future studies.
- Few predictive validity studies attempt to control confounding variables. This should be an area deserving of more methodological attention in future studies.

# 9 Appendix: An Overview of the SPICE Trials

There has been a general concern among some researchers that existing software engineering standards lack an empirical basis demonstrating that they indeed represent "good" practices. For instance, it has been noted that [141] "standards have codified approaches whose effectiveness has not been rigorously and scientifically demonstrated. Rather, we have too often relied on anecdote, 'gut feeling,' the opinions of experts, or even flawed research", and [140] "many corporate, national and international standards are based on conventional wisdom [as opposed to empirical evidence]". Similar arguments are made in [72][73][74].

Unique among software engineering standardization efforts, the developers of ISO/IEC 15504 deliberately initiated an international effort to empirically evaluate ISO/IEC 15504. This effort is known as the SPICE Trials [84][127][155].

The SPICE Trials have been divided into three broad phases to coincide with the stages that the ISO/IEC 15504 document was expected to go through on its path to international standardization. The analyses presented in this chapter come from phase 2 of the SPICE Trials.

During the trials, organizations contribute their assessment ratings data to an international trials database located in Australia, and also fill up a series of questionnaires after each assessment. The questionnaires collect information about the organization and about the assessment. There is a network of SPICE Trials co-ordinators around the world who interact directly with the assessors and the organizations conducting the assessments. This interaction involves ensuring that assessors are qualified, making questionnaires available, answering queries about the questionnaires, and following up to ensure the timely collection of data.

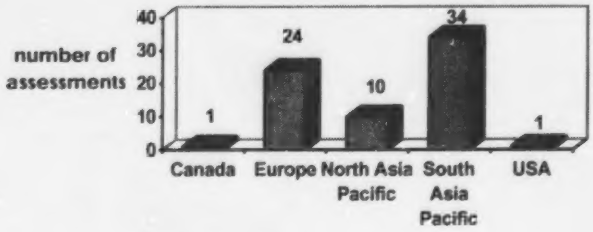


**Figure 21:** Distribution of assessments by region.

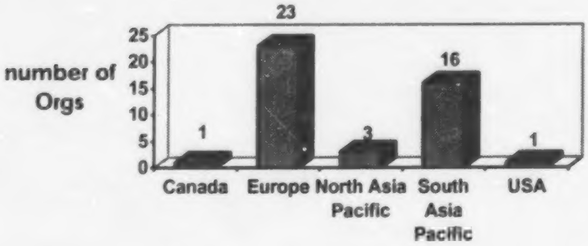


**Figure 22:** Distribution of assessed organizations by region.

A total of 70 assessments had been conducted within the context of the trials (phase 2). The distribution of assessments by region is given in Figure 21.[42] In total 691 process instances were assessed. Since more than one assessment may have occurred in a particular organization (e.g., multiple assessments each one looking at a different set of processes), a total of 44 organizations were assessed. Their distribution by region is given in Figure 22.

Given that an assessor can participate in more than one assessment, the number of assessors is smaller than the total number of assessments. In total, 40 different lead assessors took part.

---

[42] Within the SPICE Trials, assessments are coordinated within each of the five regions shown in the figures above.

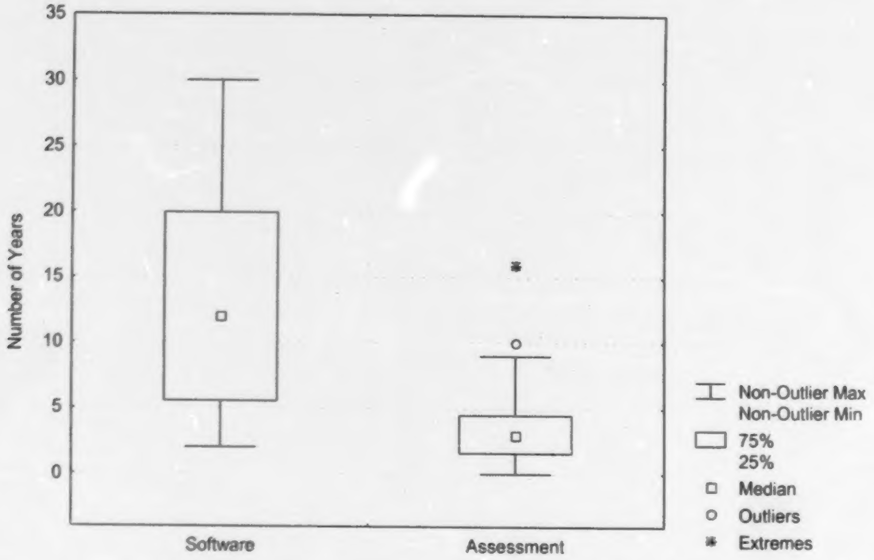


**Figure 23**: Software engineering and assessment experience of the (lead) assessors.

The variation in the number of years of software engineering experience and assessment experience of the assessors is shown in Figure 23. The median experience in software engineering is 12 years, with a maximum of 30 years experience. The median experience in assessments is 3 years, indicating a non-trivial background in assessments.

The median number of assessments performed in the past by the assessors is 6, and the median number of 15504-based assessments is 2. This indicates that, in general, assessors had a good amount of experience with software process assessments.

# 10 Appendix: A Reanalysis of the AFIT Study Data

The authors provide a detailed report of a predictive validity study of the SW-CMM maturity rating [77][120][121] using historic Air Force contracts data. Specifically, they looked at the relationship between SW-CMM maturity ratings and the ability to meet budget and schedule targets. The advantage of the data set that they use is that all contractors have to report project cost and schedule data in a consistent format, ensuring comparability of data across projects. The Air Force product centers from which data were collected were the Aeronautical Systems Center at Wright-Patterson Sir Force base, and the Electronic Systems Center at Hanscom Air Force Base.

The authors used the following criteria to select projects:

- Software specific cost and schedule data were tracked according to Cost/Schedule Control System Criteria (C/SCSC) guidelines
- The contractors were rated against the SW-CMM
- The relevance of the cost and schedule data to the SW-CMM rating could be established.

The unit of observation in this study was a software-specific WBS item in the contract (referred to as a project). Therefore, there may be more than one software project per contract. Furthermore, if a project is assessed more than once, then each assessment constitutes an observation. This is illustrated in Figure 24. Here there is one DoD contractor with two contracts A and B. For contract A there were three software specific projects, and for contract B there was only one software-specific project. The three projects in contract A were assessed twice each. In total, then, this contractor provides seven observations.[43]

Cost and schedule data were collected over the 12-month period surrounding the SW-CMM assessment date. Hence, the performance data are claimed to be *temporally relevant*. The authors also define associative relevance of the performance data to the SW-CMM rating as follows:

> **Very High Relevance** – the project under consideration was the sole project evaluated during the SW-CMM assessment
>
> **High Relevance** – the project under consideration was one of several used in obtaining the SW-CMM rating for the organization
>
> **Medium Relevance** – the project under consideration was not used to establish the SW-CMM rating, but the organization or personnel who participated in the project were also responsible for projects evaluated in the SW-CMM assessment
>
> **Low Relevance** – neither the project nor the personnel responsible for the project under consideration were used to obtain the organization's SW-CMM rating; the rating for the contractor as a whole is considered to apply to the organization responsible for the project under consideration.

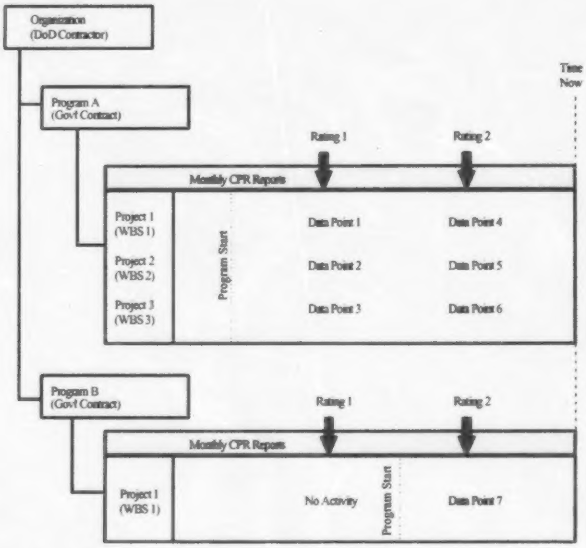


**Figure 24:** Relationship between rating period, project, and contract.

For the purposes of our analyses, we only considered the Very High Relevance and High Relevance projects. Furthermore, we excluded all assessments that were performed using the SCE method. The reason is that the focus of this paper is process improvement, and the SCE is essentially an audit rather than assessment. It is known that an assessment for the purposes of capability evaluation and for process improvement can yield different results for the same organization [12]. Finally, we also ignored all observations which the authors themselves deemed of dubious quality. We end up with data on 25 software projects, and these are included in Table 16. The authors did not perform an analysis with this particular subset, which is the subset most relevant for this review. Also note that none of the projects rated had a maturity rating greater than 3.

---

[43] Strictly speaking, under this scheme, the observations for the three projects in contract A are not independent, hence violating one of the assumptions of statistical tests.

The performance criteria were intended to measure ability to meet budget and schedule targets. These are defined as follows.

The projected rate of funds expenditure (the baseline) is expressed in the Budgeted Cost of Work Scheduled (BCWS). The Budgeted Cost of Work Performed (BCWP) represents the earned value of the work performed, and is an estimate of the work completed in dollars. The difference between the BCWS and BCWP is the schedule variance expressed in dollars, and captures the amount of work which was scheduled but not performed. The Actual Cost of Work Performed (ACWP) is the sum of funds actually expended in the accomplishment of the planned work tasks. Cost variance is the difference between what the project was expected to cost (BCWP) and what the project actually cost (ACWP).

Two indices were then defined. The Schedule Performance Index (SPI) was defined as:

$$SPI=\frac{BCWP}{BCWS} \qquad \textbf{Eqn. 3}$$

An SPI value less than 1 implies that for every dollar of work scheduled, less than one dollar has been earned – a schedule overrun. An SPI of more than 1 implies that for each dollar of work scheduled, more than one dollar of work has been earned – a schedule underrun. An SPI of 1 implies that the project was on-target. Similarly for cost we have the Cost Performance Index (CPI):

$$CPI=\frac{BCWP}{ACWP} \qquad \textbf{Eqn. 4}$$

A CPI value less than 1 indicates a cost overrun, a value greater than 1 indicates a cost underrun, and a value of 1 indicates an on-target condition.

|  | CMM Rating | SPI | CPI | SPIDEV | CPIDEV |
|---|---|---|---|---|---|
| 1 | 3.000 | .954 | 1.03526 | .04626 | .03526 |
| 2 | 3.000 | 1.024 | 1.11050 | .02420 | .11050 |
| 3 | 3.000 | 1.007 | 1.06107 | .00725 | .06107 |
| 4 | 3.000 | .934 | 1.04982 | .06646 | .04982 |
| 5 | 1.000 | 1.056 | .84981 | .05599 | .15019 |
| 6 | 3.000 | 1.026 | 1.12719 | .02618 | .12719 |
| 7 | 3.000 | .987 | .98222 | .01318 | .01778 |
| 8 | 2.000 | .966 | .38626 | .03404 | .61374 |
| 9 | 1.000 | 1.868 | .20188 | .86762 | .79812 |
| 10 | 2.000 | 1.077 | .34957 | .07744 | .65043 |
| 11 | 2.000 | 1.047 | .83683 | .04736 | .16317 |
| 12 | 3.000 | 1.000 | .96737 | 0.00000 | .03263 |
| 13 | 3.000 | 1.172 | .79640 | .17241 | .20360 |
| 14 | 3.000 | 1.086 | .98556 | .08641 | .01444 |
| 15 | 3.000 | 1.092 | .95995 | .09190 | .04005 |
| 16 | 3.000 | 1.221 | .79240 | .22072 | .20760 |
| 17 | 2.000 | .973 | .86808 | .02737 | .13192 |
| 18 | 2.000 | 1.000 | 1.21799 | 0.00000 | .21799 |
| 19 | 2.000 | .908 | 1.45455 | .09220 | .45455 |
| 20 | 2.000 | .904 | 1.11139 | .09572 | .11139 |
| 21 | 2.000 | .915 | 1.75556 | .08494 | .75556 |
| 22 | 2.000 | .988 | 2.05063 | .01220 | 1.05063 |
| 23 | 2.000 | .919 | .77919 | .08144 | .22081 |
| 24 | 2.000 | .973 | 1.14790 | .02713 | .14790 |
| 25 | 1.000 | .551 | .23626 | .44935 | .76374 |

**Table 16:** Subset of the AFIT data used for the analysis presented in this paper.

The hypothesis that the authors were testing was explicitly stated as [77] “Given that the goal of any project is to meet the target budget and schedule, an organization’s success can be measured by evaluating the CPI and SPI of a particular project. The closer the CPI and SPI are to the value of 1.00, the more successful the project can be considered, at least in terms of cost and schedule. Thus, it is reasonable to expect that as an organization’s process matures, its success or ability to consistently meet target budgets and schedules will increase”. In fact, this is the general SW-CMM hypothesis stated as “As maturity increases, the difference between targeted results and actual results decreases across projects”.

The authors then proceed to test this hypothesis using the Kruskal-Wallis test [152]. This allows the authors to test the null hypothesis that there is no difference (in the medians) among projects in each of the groups, where each group is characterized by its maturity rating (there were three groups with maturity levels 1, 2, and 3). The alternative hypothesis is that there is a difference among the groups, which would be the case if the above hypothesis was true.

However, the actual hypothesis that one wants to investigate is not whether there is any difference among the groups, but that there is better improvement as one moves from level 1 to level 3. Therefore, the alternative hypothesis is an ordered one. A better statistical test to use is therefore the Jonckheere-Terpstra (JT) test [97], which is more powerful than the Kruskal-Wallis test for ordered alternatives.

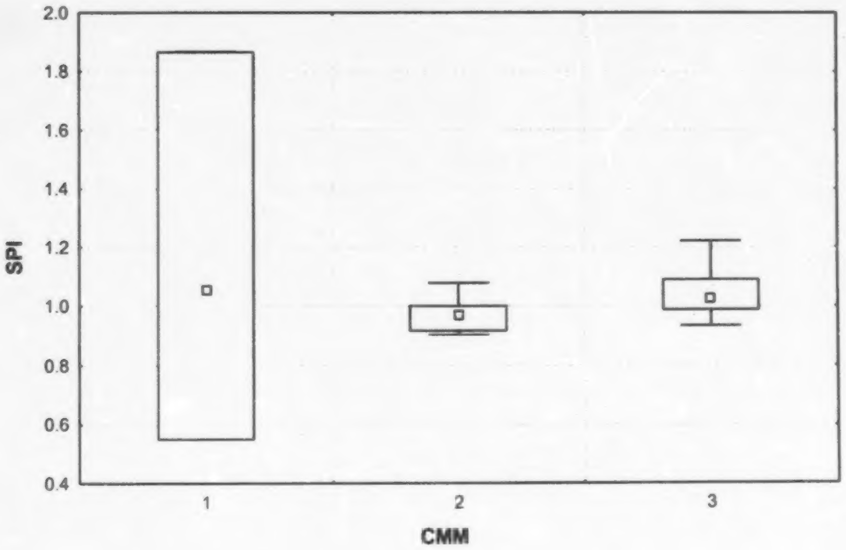


**Figure 25:** SPI values for the three levels. The J* is 1.446.

The data for the SPI performance measure are shown in Figure 25. It is seen that the median value for level 1 projects is slightly above 1 but has a large variance. It drops below 1 for level 2 projects, and goes back again above one for level 3. The JT test produces a J* value of 1.446 which has a one-sided asymptotic p-value of 0.074. This indicates that indeed there is an ordered association between SPI and maturity levels at an alpha level of 0.1.

The data for CPI are shown in Figure 26. As can be seen there is a dramatic shift from gross cost overruns for level 1 to almost meeting cost targets at levels 2 and 3. The JT test, however, produced a value of J* of 0.9554, which has an asymptotic one-sided p-value of 0.1697, indicating lack of statistical significance at an alpha level of 0.1.

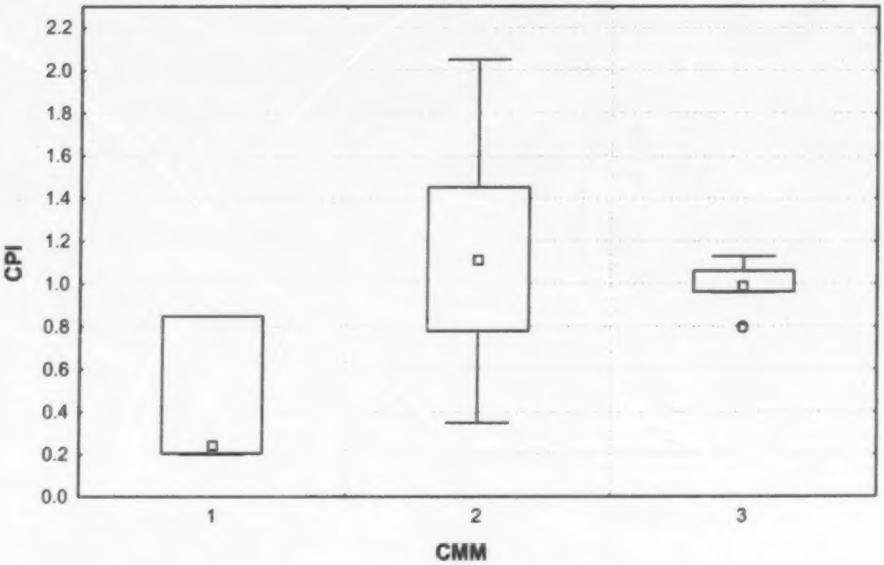


**Figure 26:** CPI values for the three levels. The J* is 0.9554.

The above results would seem to indicate that there is a relationship between meeting schedule targets and SW-CMM maturity ratings, but not with meeting cost targets.

However, a careful examination of the hypothesis that is being stated would indicate that we are actually testing the wrong hypothesis quantitatively. The hypothesis that was tested above was whether higher level projects tend to underrun their costs and schedules targets. Thus, if level 3 projects dramatically underrun costs and schedule targets then the results would be overwhelming good using the above approach. The hypothesis that we want to test is whether higher maturity is associated with meeting schedule and cost targets, not underrun them (i.e., SPI and CPI closer to one not greater than one). It is known in the literature that both overrunning targets and underrunning targets is undesirable

A proper test of the SW-CMM hypothesis as stated would therefore be to use the JT test with the following performance measures:

$$SPIDEV = \left|\frac{BCWP - BCWS}{BCWS}\right| \qquad \textbf{Eqn. 5}$$

and

$$CPIDEV = \left|\frac{BCWP - ACWP}{ACWP}\right| \qquad \textbf{Eqn. 6}$$

These deviation indices will increase as one deviates from target, over or under, and would provide the correct test of the hypothesis.

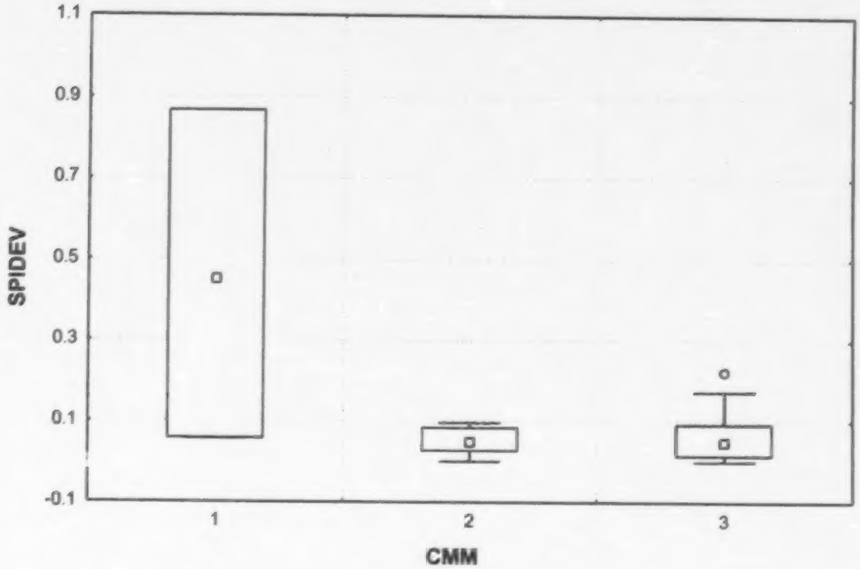


**Figure 27:** SPIDEV values for the three levels. The J* is –1.291.

Figure 27 shows the data expressed in terms of SPIDEV. Here we see a clear decrease in deviation from target as maturity level increases. The J* value of –1.291 has an asymptotic one-sided p-value of 0.0983, which is statistically significant at an alpha level of 0.1

Similarly, the results for the CPIDEV variable are illustrated in Figure 28. The J* value of –3.744 is highly statistically significant with an asymptotic one-sided p-value of 0.0001.

Therefore, our reanalysis of the AFIT data made two methodological contributions. First, we used the JT test which is known to be more powerful for ordered alternatives than the K-W test used by the authors. Second, we argued that the original analysis was not testing the authors'own hypothesis nor the SW-CMM hypothesis, and therefore we modified the performance measures to actually measure deviations from targets and used that directly.

Our results indicate that there is an association between higher capability level ratings and the ability to meet schedule and cost targets.

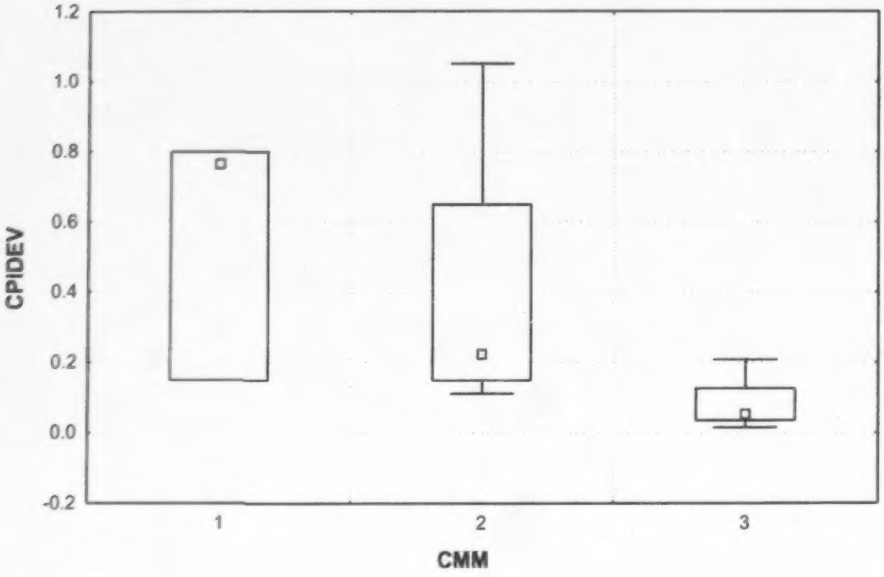


**Figure 28:** CPIDEV values for the three levels. The J* is –3.744.

# 11 Acknowledgements

We wish to thank Will Hayes and Shadia El Egazzar for providing us with thorough reviews of earlier versions of this chapter.

# 12 References

[1] N. Ahituv and S. Neumann: *Principles of Information Systems for Management*, W. Brown, 1982.

[2] M. Allen and W. Yen: *Introduction to Measurement Theory*. Brooks/Cole Publishing Company, 1979.

[3] S. Alter: *Decision support Systems: Current Practices and Continuing Challenges* Addison-Wesley, 1980.

[4] K. Amoako-Gyampah and K. White: "User Involvement and User Satisfaction: An Exploratory Contingency Model". In *Information and Management*, 25:1-10, 1993.

[5] J. Bach: "Enough About Process: What we Need Are Heroes." In *IEEE Software*, 12(2):96-98, February 1995.

[6] J. Baroudi, M. Olson, and B. Ives: "An Empirical Study of the Impact of User Involvement on System Usage and Information Satisfaction". In *Communications of the ACM*, 29(3):232-238, 1986.

[7] I. Benbasat, A. Dexter, and R. Mantha: "Impact of Organizational Maturity on Information System Skill Needs". In *MIS Quarterly*, 4(1):21-34, 1980.

[8] I. Benbasat, A. Dexter, D. Drury, and R. Goldstein: "A Critique of the Stage Hypothesis: Theory and Empirical Evidence". In *Communications of the ACM*, 27(5):476-485, 1984.

[9] I. Benbasat and R. Zmud: "Empirical Research in Information Systems: The Practice of Relevance". In *MIS Quarterly*, 23(1):3-16, March 1999.

[10] E. Bennett, R. Alpert, and A. Goldstein: "Communications Through Limited Response Questioning". In *Public Opinion Quarterly*, 18:303-308, 1954.

[11] S. Benno and D. Frailey: "Software Process Improvement in DSEG: 1989-1995". In *Texas Instruments Technical Journal*, 12(2):20-28, March-April 1995.

[12] J. Besselman, P. Byrnes, C. Lin, M. Paulk, and R. Puranik: "Software Capability Evaluations: Experiences from the Field". In *SEI Technical Review*, 1993.

[13] T. Bollinger and C. McGowan: "A Critical Look at Software Capability Evaluations". In *IEEE Software*, pages 25-41, July 1991.

[14] Bootstrap Project Team: "Bootstrap: Europe's Assessment Method". In *IEEE Software*, pages 93-95, May 1993.

[15] J. Brodman and D. Johnson: "What Small Businesses and Small Organizations Say about the CMM". In *Proceedings of the 16th International Conference on Software Engineering*, pages 331-340, 1994.

[16] J. Brodman and D. Johnson: "Return on Investment (ROI) from Software Process Improvement as Measured by US Industry". In *Software Process: Improvement and Practice*, Pilot Issue, John Wiley & Sons, 1995.

[17] J. Brodman and D. Johnson: "Return on Investment from Software Process Improvement as Measured by U.S. Industry". In *Crosstalk*, 9(4):23-29, April 1996.

[18] C. Buchman: "Software Process Improvement at AlliedSignal Aerospace". In *Proceedings of the 29th Annual Hawaii International Conference on Systems Science, Vol. 1: Software Technology and Architecture*, pages 673-680, 1996.

[19] F. Budlong and J. Peterson: "Software Metrics Capability Evaluation Guide". The Software Technology Support Center, Ogden Air Logistics Center, Hill Air Force Base, 1995.

[20] I. Burnstein, T. Suwannasart, and C. Carlson: "Developing a Testing Maturity Model: Part I". In *Crosstalk*, pages 21-24, August 1996.

[21] I. Burnstein, T. Suwannasart, and C. Carlson: "Developing a Testing Maturity Model: Part II". In *Crosstalk*, pages 1926-24, September 1996.

[22] K. Butler: "The Economic Benefits of Software Process Improvement". In *Crosstalk*, 8(7):14-17, July 1995.

[23] D. Card: "Understanding Process Improvement". In *IEEE Software*, pages 102-103, July 1991.

[24] D. Card: "Capability Evaluations Rated Highly Variable". In *IEEE Software*, pages 105-106, September 1992.

[25] B. Clark: *The Effects of Software Process Maturity on Software Development Effort*. PhD Thesis, University of Southern California, April 1997.

[26] E. Carmines and R. Zeller: *Reliability and Validity Assessment*, Sage Publications, Beverly Hills, 1979.

[27] F. Coallier, J. Mayrand, and B. Lague: "Risk Management in Software Product Procurement". In K. El Emam and N. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[28] J. Cohen: "A Coefficient of Agreement for Nominal Scales". In *Educational and Psychological Measurement*, XX(1):37-46, 1960.

[29] J. Cohen: "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Agreement or Partial Credit". In *Psychological Bulletin*, 70:213-220, 1968.

[30] M. Craigmyle and I. Fletcher: "Improving IT Effectiveness through Software Process Assessment". In *Software Quality Journal*, 2:257-264, 1993.

[31] L. Cronbach: "Coefficient Alpha and the Internal Structure of Tests". In *Psychomterika*, 16(3):297-334, 1951.

[32] L. Cronbach, G. Gleser, H. Nanda, and N. Rajaratnam: *The Dependability of Behavioral Measurements: Theory of Generalizability of Scores and Profiles*, John Wiley, 1972.

[33] L. Cronbach: "Test Validation". In R. Thorndike (ed.): *Educational Measurement*, American Council on Education, 1971.

[34] B. Curtis, H. Krasner, and N. Iscoe: "A Field Study of the Software Design Process for Large Systems". In *Communications of the ACM*, 31(11):1268-1286, November 1988.

[35] B. Curtis: "The Factor Structure of the CMM and Other Latent Issues". Paper presented at the *Empirical Studies of Programmers: Sixth Workshop*, Washington DC, 1996.

[36] B. Curtis, W. Hefley, S. Miller, and M. Konrad: "The People Capability Maturity Model for Improving the Software Workforce". In K. El Emam and N. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[37] G. Dadoun: "ISO 9000: A Requirement for Doing Business". In *Proceedings of the CAS (Centre for Advanced Studies) Conference*, IBM Canada Ltd., pages 433-437, 1992.

[38] F. Davis: "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In *MIS Quarterly*, pages 319-340, 1989.

[39] C. Deephouse, D. Goldenson, M. Kellner, and T. Mukhopadhyay: "The Effects of Software Processes on Meeting Targets and Quality". In *Proceedings of the Hawaiian International Conference on Systems Sciences*, vol. 4, pages 710-719, January 1995.

[40] R. Dion: "Elements of a Process Improvement program". In *IEEE Software*, 9(4):83-85, July 1992.

[41] R. Dion: "Process Improvement and the Corporate Balance Sheet". In *IEEE Software*, 10(4):28-35, July 1993.

[42] D. Drehmer and S. Dekleva: "Measuring Software Engineering Maturity: A Rasch Calibration". In *Proceedings of the International Conference on Information Systems*, pages 191-202, 1993.

[43] D. Drew: "Tailoring the Software Engineering Institute's (SEI) Capability Maturity Model (CMM) to a Software Sustaining Engineering Organization". In *Proceedings of the International Conference on Software Maintenance*, pages 137-144, 1992.

[44] J-N Drouin: "Introduction to SPICE". In K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*, IEEE CS Press, 1998.

[45] D. Drury: "An Empirical Assessment of the Stages of Data Processing Growth". In *MIS Quarterly*, 7(2):59-70, 1983.

[46] D. Dunaway: "CMM-Based Appraisal for Internal Process Improvement (CBA IPI) Lead Assessor's Guide."Software Engineering Institute, Handbook CMU/SEI-96-HB-003, 1996.

[47] D. Dunnaway, D. Goldenson, I. Monarch, and D. White: "How Well is CBA IPI Working ? User Feedback". In *Proceedings of the 1998 Software Engineering Process Group Conference*, 1998.

[48] D. Dunaway and S. Masters: *CMM-Based Appraisal for Internal Process Improvement (CBA IPI): Method Description.* Software Engineering Institute, Technical Report CMU/SEI-96-TR-007, 1996.

[49] K. Dymond: "Essence and Accidents in SEI-Style Assessments or Maybe This Time the Voice of the Engineer will be Heard". In K. El Emam and N. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[50] P. Ein-Dor and E. Segev: *Managing Management Information Systems*, Lexington, 1978.

[51] K. El Emam and N. H. Madhavji: "A Method for Instrumenting Software Evolution Processes and An Example Application". In *Notes From The International Workshop on Software Evolution, Processes, and Measurements*, Technical Report #94-04 NT, Software Engineering Test Lab, Department of Computer Science, University of Idaho, 1994.

[52] K. El Emam and D. R. Goldenson: "SPICE: An Empiricist's Perspective". In *Proceedings of the Second IEEE International Software Engineering Standards Symposium*, pages 84-97, August 1995.

[53] K. El Emam and N. H. Madhavji: "The Reliability of Measuring Organizational Maturity". In *Software Process: Improvement and Practice*, 1(1):3-25, 1995.

[54] K. El Emam and N. H. Madhavji: "A Field Study of Requirements Engineering Practices in Information Systems Development". In *Proceedings of the Second IEEE International Symposium on Requirements Engineering*, pages 68-80, 1995.

[55] K. El Emam and N. H. Madhavji: "Measuring the Success of Requirements Engineering Processes". In *Proceedings of the Second IEEE International Symposium on Requirements Engineering*, pages 204-211, 1995.

[56] K. El Emam and D. Goldenson: "An Empirical Evaluation of the Prospective International SPICE Standard". In *Software Process – Improvement and Practice*, 2:123-148, 1996.

[57] K. El Emam, L. Briand, and B. Smith: "Assessor Agreement in Rating SPICE Processes". In *Software Process Improvement and Practice Journal*, 2(4):291-306, 1996.

[58] K. El Emam, D. Goldenson, L. Briand, and P. Marshall: "Interrater Agreement in SPICE-Based Assessments: Some Preliminary Results". In *Proceedings of the International Conference on the Software Process*, pages 149-156, 1996.

[59] K. El Emam, B. Smith, and P. Fusaro: "Modeling the Reliability of SPICE Based Assessments". In *Proceedings of the Third IEEE International Software Engineering Standards Symposium*, pages 69-82, 1997.

[60] K. El Emam: "The Internal Consistency of the ISO/IEC 15504 Software Process Capability Scale". In *Proceedings of the 5th International Symposium on Software Metrics*, pages 72-81, IEEE CS Press, 1998.

[61] K. El Emam, J-M Simon, S. Rousseau, and E. Jacquet: "Cost Implications of Interrater Agreement for Software Process Assessments". In *Proceedings of the 5th International Symposium on Software Metrics*, pages 38-51, IEEE CS Press, 1998.

[62] K. El Emam and P. Marshall: "Interrater Agreement in Assessment Ratings". In K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press, 1998.

[63] K. El Emam and L. Briand: "Costs and Benefits of Software Process Improvement". In R. Messnarz and C. Tully (eds.): *Better Software Practice for Business Benefit: Principles and Experience*. IEEE CS Press, (to appear) 1999.

[64] K. El Emam, J-N Drouin, W. Melo: *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press, 1998.

[65] K. El Emam, B. Smith, P. Fusaro: "Success Factors and Barriers for Software Process Improvement: An Empirical Study". In R. Messnarz and C. Tully (eds.): *Better Software Practice for Business Benefit: Principles and Experiences*, IEEE CS Press, 1999.

[66] K. El Emam and A. Birk: "Validating the ISO/IEC 15504 Measures of Software Development Process Capability". To appear in the *Journal of Systems and Software*.

[67] K. El Emam and A. Birk: "Validating the ISO/IEC 15504 Measures of Software Requirements Analysis Process Capability". To appear in *IEEE Transactions on Software Engineering*.

[68] K. El Emam: "Benchmarking Kappa: Interrater agreement in software process assessments". To appear in *Empirical Software Engineering: An International Journal*, Kluwer Academic Publishers.

[69] K. El Emam and I. Garro: "Estimating the Extent of Standards Use: The Case of ISO/IEC 15504". Under revision for the *Journal of Systems and Software*.

[70] K. El Emam and N. H. Madhavji: "The Impact of User Capability on Requirements Engineering Success". Submitted for publication.

[71] N. Fenton: "Objectives and Context of Measurement/Experimentation". In H. D. Rombach, V. Basili, and R. Selby (eds.): *Experimental Software Engineering Issues: Critical Assessment and Future Directions*, Springer-Verlag, 1993.

[72] N. Fenton, B. Littlewood, and S. Page: "Evaluating Software Engineering Standards and Methods". In R. Thayer and A. McGettrick (eds.): *Software Engineering: A European Perspective*, IEEE CS Press, 1993.

[73] N. Fenton and S. Page: "Towards the Evaluation of Software Engineering Standards". In *Proceedings of the Software Engineering Standards Symposium*, pages 100-107, 1993.

[74] N. Fenton, S-L Pfleeger, S. Page, and J. Thornton: "The SMARTIE Standards Evaluation Methodology". *Technical Report* (available from the Centre for Software Reliability, City University, UK), 1994.

[75] P. Finlay and A. Mitcheil: "Perceptions of the Benefits from the Introduction of CASE: An Empirical Study". In *MIS Quarterly*, pages 353-370, December 1994.

[76] J. Fleiss: *Statistical Methods for Rates and Proportions*, John Wiley & Sons, 1981.

[77] R. Flowe and J. Thordahl: *A Correlational Study of the SEI's Capability Maturity Model and Software Development Performance in DoD Contracts*. MSc Thesis, The Air Force Institute of Technology, 1994.

[78] P. Fusaro, K. El Emam, and B. Smith: "Evaluating the Interrater Agreement of Process Capability Ratings". In *Proceedings of the Fourth International Software Metrics Symposium*, pages 2-11, 1997.

[79] P. Fusaro, K. El Emam, and B. Smith: "The Internal Consistencies of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension". In *Empirical Software Engineering: An International Journal*, 3:179-201, Kluwer Academic Publishers, 1997.

[80] D. Galletta and A. Ledrer: "Some Cautions on the Measurement of User Information Satisfaction". In *Decision Sciences*, 20:419-438, 1989.

[81] C. Gibson and R. Nolan: "Behavioral and Organizational Issues in the Stages of Managing the Computer Resource". In R. Nolan (ed.): *Managing the Data Resource Function*, West Publishing Company, 1974.

[82] C. Gibson and R. Nolan: "Managing the Four Stages of EDP Growth". In *Harvard Business Review*, pages 76-88, 1974.

[83] D. R. Goldenson and J. D. Herbsleb: "After the Appraisal: A Systematic Survey of Process Improvement, its Benefits, and Factors that Influence Success". Technical Report, CMU/SEI-95-TR-009, Software Engineering Institute, 1995.

[84] D. R. Goldenson and K. El Emam: "The international SPICE trials: Project description and initial results". In *Proceedings of the 8th Software Engineering Process Group Conference*, May 1996.

[85] D. Goldenson, K. El Emam, J. Herbsleb, and C. Deephouse: "Empirical Studies of Software Process Assessment Methods". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[86] R. Goldstein and I. McCririck: "The Stage Hypothesis and Data Administration: Some Contradictory Evidence". In *Proceedings of the 2nd International Conference on Information Systems*, pages 309-324, 1981.

[87] E. Gray and W. Smith: "On the Limitations of Software Process Assessment and the Recognition of a Required Re-Orientation for Global Process Improvement". In *Software Quality Journal*, 7:21-34, 1998.

[88] J-F Gregoire and F. Lustman: "The Stage Hypothesis Revisited: An EDP Professionals' Point of View". In *Information and Management*, 24:237-245, 1993.

[89] V. Haase, R. Messnarz, G. Koch. H. Kugler, and P. Decrinis: "Bootstrap: Fine-Tuning Process Assessment". In *IEEE Software*, pages 25-35, July 1994.

[90] D. Hartmann: "Considerations in the Choice of Interobserver Reliability Estimates". In *Journal of Applied Behavior Analysis*, 10(1):103-116, 1977.

[91] W. Hayes and D. Zubrow: "Moving On Up: Data and Experience Doing CMM-Based Process Improvement". Technical Report CMU/SEI-95-TR-008, Software Engineering Institute, 1995.

[92] J. Herbsleb, A. Carleton, J. Rozum, J. Siegel, and D. Zubrow: "Benefits of CMM-Based Software Process Improvement: Initial Results". Technical Report, CMU-SEI-94-TR-13, Software Engineering Institute, 1994.

[93] J. Herbsleb, D. Zubrow, D. Goldenson, W. Hayes, and M. Paulk: "Software Quality and the Capability Maturity Model". In *Communications of the ACM*, 40(6):30-40, 1997.

[94] J. Herbsleb: "Hard Problems and Hard Science: On the Practical Limits of Experimentation". In *IEEE TCSE Software Process Newsletter*, No. 11, pages 18-21, 1998.

[95] J. Herbsleb and R. Grinter: "Conceptual Simplicity Meets Organizational Complexity: Case Study of a Corporate Metrics Program". In *Proceedings of the 20th International Conference on Software Engineering*, pages 271-280, 1998.

[96] A. Hersh: "Where's the Return on Process Improvement ?". In *IEEE Software*, page 12, July 1993.

[97] M. Hollander and D. Wolfe: *Nonparametric Statistical Methods*. Wiley, 1999.

[98] A. Huebner: "ISO 9000 Implementation in Software Development of IBM Germany". Paper presented at SDC 1992, IBM, Application development Germany, May 7 1992.

[99] S. Huff, M. Munro, and B. Martin: "Growth Stages of End User Computing". In *Communications of the ACM*, 31(5):542-550, 1988.

[100] W. Humphrey: "Characterizing the Software Process: A Maturity Framework". In *IEEE Software*, pages 73-79, March 1988.

[101] W. Humphrey and W. Sweet: "A Method for Assessing the Software Engineering Capability of Contractors". Technical Report CMU/SEI-87-TR-23, Software Engineering Institute, 1987.

[102] W. Humphrey, T. Snyder, and R. Willis: "Software Process Improvement at Hughes Aircraft". In *IEEE Software*, pages 11-23, July 1991.

[103] W. Humphrey and B. Curtis: "Comments on 'A Critical Look'". In *IEEE Software*, pages 42-46, July 1991.

[104] ISO/IEC TR 15504: *Information Technology – Software Process Assessment*, 1998. (parts 1-9; part 5 was published in 1999). Available from http://www.iese.fhg.de/SPICE.

[105] B. Ives, M. Olson, and J. Baroudi: "The Measurement of User Information Satisfaction". In *Communications of the ACM*, 26(10):785-793, 1983.

[106] Japan SC7 WG10 SPICE Committee: "Report of Japanese Trials Process Assessment by SPICE Method". *A SPICE Project Report*, 1994.

[107] C. Jones: "The Pragmatics of Software Process Improvements". In *Software Process Newsletter*, IEEE Computer Society TCSE, No. 5, pages 1-4, Winter 1996. . (available at http://www-se.cs.mcgill.ca/process/spn.html)

[108] C. Jones: "The Economics of Software Process Improvements". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[109] C. Jones: *Assessment and Control of Software Risks*. Prentice-Hall, 1994.

[110] C. Jones: "Gaps in SEI Programs." In *Software Development*, 3(3):41-48, March 1995.

[111] J. Kim and C. Mueller: *Factor Analysis: Statistical Methods and Practical Issues*. Sage Publications, 1978.

[112] J. King and K. Kraemer: "Evolution and Organizational Information Systems: An Assessment of Nolan's Stage Model". In *Communications of the ACM*, 27(5):466-475, 1984.

[113] P. Keen and M. Scott Morton: *Decision Support Systems: An Organizational Perspective*. Addison-Wesley, 1978.

[114] F. Kerlinger: *Foundations of Behavioral Research*. Holt, Rinehart, and Winston, 1986.

[115] M. Khurana and K. El Emam: "Assessment Experience and the Reliability of Assessments". In *Software Process Newsletter*, IEEE Technical Council on Software Engineering, No. 12, Spring 1998.

[116] D. Kitson and S. Masters: "An Analysis of SEI Software Process Assessment Results: 1987-1991". In *Proceedings of the International Conference on Software Engineering*, pages 68-77, 1993.

[117] D. Klein: "If You Get Straight A's, You Must be Intelligent – Respecting the Intent of the Capability Maturity Model". In *Crosstalk*, pages 22-23, February 1998.

[118] H. Krasner: "The Payoff for Software Process Improvement: What it is and How to Get it". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[119] M. Krishnan and M. Kellner: "Measuring Process Consistency: Implications for Reducing Software Defects". March 1998. Submitted for Publication.

[120] P. Lawlis, R. Flowe, and J. Thordahl: "A Correlational study of the CMM and Software Development Performance". In *Crosstalk*, pages 21-25, September 1995.

[121] P. Lawlis, R. Flowe, and J. Thordahl: "A Correlational Study of the CMM and Software Development Performance". In *Software Process Newsletter*, IEEE TCSE, No. 7, pages 1-5, Fall 1996. (available at http://www-se.cs.mcgill.ca/process/spn.html)

[122] L. Lebsanft: "Bootstrap: Experiences with Europe's Software Process Assessment and Improvement Method". In *Software Process Newsletter*, IEEE Computer Society, No. 5, pages 6-10, Winter 1996. (available at http://www-se.cs.mcgill.ca/process/spn.html)

[123] J. Lee and S. Kim: "The Relationship between Procedural Formalization in MIS Development and MIS Success". In *Information and Management*, 22:89-111, 1992.

[124] W. Lipke and K. Butler: "Software Process Improvement: A Success Story". In *Crosstalk*, 5(9):29-39, September 1992.

[125] F. Lord and M. Novick: *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

[126] H. Lucas and J. Sutton: "The Stage Hypothesis and the S-Curve: Some Contradictory Evidence". In *Communications of the ACM*, 20(4):254-259, 1977.

[127] F. Maclennan, G. Ostrolenk, and M. Tobin: "Introduction to the SPICE Trials". In K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*, IEEE CS Press, 1998.

[128] S. Masters and C. Bothwell: *CMM Appraisal Framework – Version 1.0.* Software Engineering Institute, Technical Report CMU/SEI-TR-95-001, 1995.

[129] B. McFeeley: "IDEAL: A User's Guide for Software Process Improvement." Software Engineering Institute, Technical Report CMU/SEI-96-HB-001, February 1996..

[130] F. McGarry, S. Burke, and B. Decker: "Measuring the Impacts Individual Process Maturity Attributes Have on Software Projects". In *Proceedings of the 5th International Software Metrics Symposium*, pages 52-60, 1998.

[131] T. McGibbon: "A Business Case for Software Process Improvement Revised: Measuring Return on Investment from Software Engineering and Management". DoD Data & Analysis Center for Software (DACS), Technical Report, September 1999 (available from <http://www.dacs.dtic.mil/techs/roispi2/roispi2.pdf>).

[132] J. Neter, W. Wasserman, and M. Kunter: *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs*, Irwin, 1990.

[133] R. Nolan: "Managing the Computer Resource: A Stage Hypothesis". In *Communications of the ACM*, 16(7):399-405, 1973.

[134] R. Nolan: "Thoughts About the Fifth Stage". In *Database*, 7(2):4-10, 1975.

[135] R. Nolan: "Managing the Crisis in Data Processing". In *Harvard Business Review*, pages 115-126, March/April 1979.

[136] J. Nunnally and I. Bernstein: *Psychometric Theory*. McGraw Hill, 1994.

[137] M. Paulk and M. Konrad: "Measuring Process Capability versus Organizational Process Maturity". In *Proceedings of the 4th International Conference on Software Quality*, October 1994.

[138] M. Paulk: "The Evolution of the SEI's Capability Maturity Model for Software". In *Software Process – Improvement and Practice*, Pilot Issue, Pages 1-15, 1995.

[139] M. Paulk, C. Weber, and M-B Chrissis: "The Capability Maturity Model for Software". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[140] S-L Pfleeger: "The Language of Case Studies and Formal Experiments". In *Software Engineering Notes*, pages 16-20, October 1994.

[141] S-L Pfleeger, N. Fenton, and S. Page: "Evaluating Software Engineering Standards". In *IEEE Computer*, pages 71-79, September 1994.

[142] S-L Pfleeger: "Understanding and Improving Technology Transfer in Software Engineering". In *The Journal of Systems and Software*, 47:111-124, 1999.

[143] R. Radice, J. Harding, P. Munnis, and R. Phillips: "A Programming Process Study". In *IBM Systems Journal*, 24(2):91-101, 1985.

[144] S. Raghavan and D. Chand: "Diffusing Software-Engineering Methods". In *IEEE Software*, pages 81-90, July 1989.

[145] S. Rahhal: *An Effort Estimation Model for Implementing ISO 9001 in Software Organizations* Master's Thesis, School of Computer Science, McGill University, October 1995.

[146] D. Rubin: *Multiple Imputation for Nonresponse in Surveys* John Wiley & Sons, 1987.

[147] H. Rubin: "Software Process Maturity: Measuring its Impact on Productivity and Quality". In *Proceedings of the International Conference on Software Engineering* pages 468-476, 1993.

[148] D. Rugg: "Using a Capability Evaluation To Select A Contractor". In *IEEE Software*, pages 36-45, July 1993.

[149] H. Saiedian and R. Kuzara: "SEI Capability Maturity Model's Impact on Contractors". In *IEEE Computer*, pages 16-26, January 1995.

[150] W. Scott: "Reliability of Content Analysis: The Case of Nominal Scale Coding". In *Public Opinion Quarterly*, 19:321-325, 1955.

[151] V. Sethi and W. King: "Construct Measurement in Information Systems Research: An Illustration in Strategic Systems". In *Decision Sciences*, 22:455-472, 1991.

[152] S. Siegel and N. J. Castellan: *Nonparametric Statistics for the Behavioral Sciences* McGraw-Hill, 1988.

[153] J-M Simon, K. El Emam, S. Rousseau, E. Jacquet, and F. Babey: "The Reliability of ISO/IEC PDTR 15504 Assessments". In *Software Process Improvement and Practice Journal*, 3:177-188, 1997.

[154] J-M Simon: "Assessment Using SPICE: A Case Study". In K. El Emam, J-N Drouin, W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination* IEEE CS Press, 1998.

[155] B. Smith and K. El Emam: "Transitioning to phase 2 of the SPICE trials". In *Proceedings of SPICE'96*, pages 45-55, 1996.

[156] Software Engineering Institute: "A Systems Engineering Capability Maturity Model, Version 1.0". Software Engineering Institute, Handbook CMU/SEI-94-HB-04, 1994.

[157] Software Engineering Institute: *The Capability Maturity Model: Guidelines for Improving the Software Process*. Addison Wesley, 1995.

[158] I. Sommerville and P. Sawyer: *Requirements Engineering: A Good Practice Guide* John Wiley & Sons, 1997.

[159] B. Springsteen, B. Brykczynski, D. Fife, R. Meeson, and J. Norris: "Policy Assessment for the Software Process Maturity Model". Institute for DefenSe Analysis, Report IDA-D-1202, 1992.

[160] D. Stelzer and W. Mellis: "Success Factors of Organizational Change in Software Process Improvement". To appear in *Software Process – Improvement and Practice*.

[161] H. Suen and P. Lee: 'The Effects of the Use of Percentage Agreement on Behavioral Observation Reliabilities: A Reassessment'. In *Journal of Psychopathology and Behavioral Assessment* 7(3):221-234, 1985.

[162] P. Tait and I. Vessey: 'The Effect of User Involvement on System Success: A Contingency Approach'" In *MIS Quarterly*, pages 91-108, March 1988.

[163] The SPIRE Project: *The SPIRE Handbook: Better Faster Cheaper Software Development in Small Companies.* ESSI Project 23873, November 1998.

[164] G. Torkzadeh and W. Doll: 'The Test-Retest Reliability of User Involvement Instruments'. In *Information and Management*, 26:21-31, 1994.

[165] Software Engineering Laboratory: *Software Process Improvement Guidebook*. NASA/GSFC, Technical Report SEL-95-002, 1995.

[166] Staff: 'A Survey of the Surveys on the Benefits of ISO 9000'. In *Software Process, Quality & ISO 9000*, 3(11):1-5, November 1994.

[167] H. Steinen: 'Software Process Assessment and Improvement: 5 Years of Experiences with Bootstrap'. In K. El Emam and N. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[168] A. Subramanian and S. Nilakanta: 'Measurement: A Blueprint for Theory Building in MIS'. In *Information and Management*, 26:13-20, 1994.

[169] S. Weissfelner: 'ISO 9001 for Software Organizations'. In K. El Emam and N. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[170] R. Whitney, E. Nawrocki, W. Hayes, and J. Siegel: 'Interim Profile: Development and Trial of a Method to Rapidly Measure Software Engineering Maturity Status'. Technical Report, CMU/SEI-94-TR-4, Software Engineering Institute, 1994.

[171] H. Wohlwend and S. Rosenbaum: 'Software Improvements in an International Company'. In *Proceedings of the International Conference on Software Engineering*, pages 212-220, 1993.

[172] R. Zwick: "Another Look at Interrater Agreement". In *Psychological Bulletin*, 103(3):374-378, 1988.